VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA STROJNÍHO INŽENÝRSTVÍ
FACULTY OF MECHANICAL ENGINEERING

ÚSTAV MATEMATIKY
INSTITUTE OF MATHEMATICS

APROXIMACE PROSTOROVĚ DISTRIBUOVANÝCH HIERARCHICKY STRUKTUROVANÝCH DAT
APPROXIMATION OF SPATIALLY-DISTRIBUTED HIERARCHICALLY ORGANIZED DATA

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE — **Bc. Veronika Smejkalová**
AUTHOR

VEDOUCÍ PRÁCE — **Ing. Martin Pavlas, Ph.D.**
SUPERVISOR

BRNO 2018

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA STROJNÍHO INŽENÝRSTVÍ

Zadání diplomové práce

Ústav:	Ústav matematiky
Studentka:	**Bc. Veronika Smejkalová**
Studijní program:	Aplikované vědy v inženýrství
Studijní obor:	Matematické inženýrství
Vedoucí práce:	**Ing. Martin Pavlas, Ph.D.**
Akademický rok:	2017/18

Ředitel ústavu Vám v souladu se zákonem č.111/1998 o vysokých školách a se Studijním a zkušebním řádem VUT v Brně určuje následující téma diplomové práce:

Aproximace prostorově distribuovaných hierarchicky strukturovaných dat

Stručná charakteristika problematiky úkolu:

Síťové úlohy a související logistické úlohy z oblasti supply–chain management reprezentují efektivní nástroje pro plánování výrobních a zpracovatelských kapacit a dopravy mezi nimi.
Řešení reálných úloh je spojeno s potřebou identifikace, aproximace a verifikace vstupních dat tak, aby je bylo možné využít pro robustní prognózování (extrapolaci) dat o produkci odpadu v mnoha geografických bodech a více budoucích časech.
V kombinaci s existencí poměrně omezených a neúplných zdrojů historických dat se jedná o komplexní úlohu preprocessingu optimalizačních úloh, která vyžaduje vhodnou kombinaci klasických statistických přístupů regresní analýzy a zpracování časových řad, modelů celočíselné optimalizace a metod jejich řešení.
V průběhu zpracování bude práce konzultována se specialisty z ÚM (Ing. Josef Bednář, Ph.D., RNDr. Pavel Popela, Ph.D. a Ing. Jakub Kůdela).
Zadání souvisí s vývojem nástroje Justine (upi.fme.vutbr.cz/justine) a řešením projektu WtECC (Centrum kompetence pro energetické využití odpadů) a NETME Centre Plus.

Cíle diplomové práce:

– Prohloubení znalostí pokročilých statistických metod
– Modelování neurčitosti a neúplnosti dat extrapolačního modelu aplikací wait–and–see přístupu stochastické optimalizace
– Vývoj modelů pro redukci neurčitosti vhodnou aproximací v rozsáhlých hierarchických strukturách.
– Testování modelu na reálných datech v oblasti odpadového hospodářství a jejich využití v rámci komplexních úloh.

Seznam doporučené literatury:

CRESSIE, N. A. C. Statistics for Spatial Data, Revised Edition. Book Series: Wiley Series in Probability and Statistics, John Wiley & Sons, Inc., 1993.

NASH, S. et al. Linear and nonlinear programming. McGraw-Hill, 1995.

WALLACE, S. W. a A. J. KING. Modeling with Stochastic Programming. Springer Verlag, 2012.

WOLSEY, L. A. Integer programming. John Wiley and Sons, 1998.

WILLIAMS, H. P. Model Building for Mathematical Programming, Wiley and Sons, 1993.

GAMS Modelling Language Manuals, GAMS, 2015.

ZVÁRA, K. a J. ŠTEPÁN. Pravdepodobnost a matematická statistika. 2. vyd. Praha: Matfyzpress, MFF UK, 2002.

ANDEL, J. Statistické metody. 2. vyd., Praha: Matfyzpress, 1998.

Termín odevzdání diplomové práce je stanoven časovým plánem akademického roku 2017/18

V Brně, dne

L. S.

prof. RNDr. Josef Šlapal, CSc. ředitel ústavu	doc. Ing. Jaroslav Katolický, Ph.D. děkan fakulty

ABSTRAKT

Prognóza produkce odpadu je důležitou informací pro plánování v oblasti odpadového hospodářství. Historická data často disponují pouze krátkou časovou řadou, kde tradiční prognostické přístupy selhávají. Tato práce navrhuje matematický model pro odhad budoucí produkce odpadu v prostorově distribuovaných datech s hierarchickou strukturou. Přístup vychází z principů regresní analýzy se závěrečnou bilancí pro soulad hodnot agregovaných dat. Součástí modelu je volba tvaru regresní funkce modelující trend v datech. Přičemž data jsou očištěna od odlehlých pozorování, která se v databázi poměrně hojně vyskytují. Důraz je kladen na rozložení rozsáhlého modelu na podúlohy, které jsou řešeny jednotlivě a vedou ke snazší implementaci. Výstupem je komplexní výpočetní nástroj, který byl testován v rámci případové studie na datech o produkci komunálního odpadu v České republice.

KLÍČOVÁ SLOVA

prognostické modely, regresní analýza, nelineární regrese, krátká časová řada, trend v datech

ABSTRACT

The forecast of the waste production is an important information for planning in waste management. The historical data often consists of short time series, therefore traditional prognostic approaches fail. The mathematical model for forecasting of future waste production based on spatially distributed data with hierarchically structure is suggested in this thesis. The approach is based on principles of regression analysis with final balance to ensure the compliance of aggregated data values. The selection of the regression function is a part of mathematical model for high-quality description of data trend. In addition, outlier values are cleared, which occur abundantly in the database. The emphasis is on decomposition of extensive model into subtasks, which lead to a simpler implementation. The output of this thesis is tool tested within case study on municipal waste production data in the Czech Republic.

KEYWORDS

forecasting models, regression analysis, nonlinear regression, short time series, trend in data

SMEJKALOVÁ, Veronika *Aproximace prostorově distribuovaných hierarchicky strukturovaných dat*: diplomová práce. BRNO: Vysoké učení technické v Brně, Fakulta strojního inženýrství, Ústav Matematiky, 2018. 76 s. Vedoucí práce Ing. Martin Pavlas, Ph.D.

PROHLÁŠENÍ

Prohlašuji, že svou diplomovou práci na téma „Aproximace prostorově distribuovaných hierarchicky strukturovaných dat“ jsem vypracoval samostatně pod vedením vedoucího diplomové práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou všechny citovány v práci a uvedeny v seznamu literatury na konci práce.

Jako autor uvedené diplomové práce dále prohlašuji, že v souvislosti s vytvořením této diplomové práce jsem neporušil autorská práva třetích osob, zejména jsem nezasáhl nedovoleným způsobem do cizích autorských práv osobnostních a/nebo majetkových a jsem si plně vědom následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon), ve znění pozdějších předpisů, včetně možných trestněprávních důsledků vyplývajících z ustanovení části druhé, hlavy VI. díl 4 Trestního zákoníku č. 40/2009 Sb.

BRNO ....

podpis autora

PODĚKOVÁNÍ

Na tomto místě bych ráda poděkovala Ing. Martinu Pavlasovi, Ph.D. za vedení práce a Ing. Radovanu Šomplákovi, Ph.D. za trpělivost, cenné rady a celou řadu konzultací. Za poskytnutý čas dále děkuji Ing. Vlastimíru Nevrlému a Ing. Jakubu Kůdelovi. Velký dík patří rodičům, sestřičce Moničce a celé naší velké rodině za podporu při studiu. Závěrem děkuji všem spolužákům, protože s Vámi bylo studium vždy zábava.

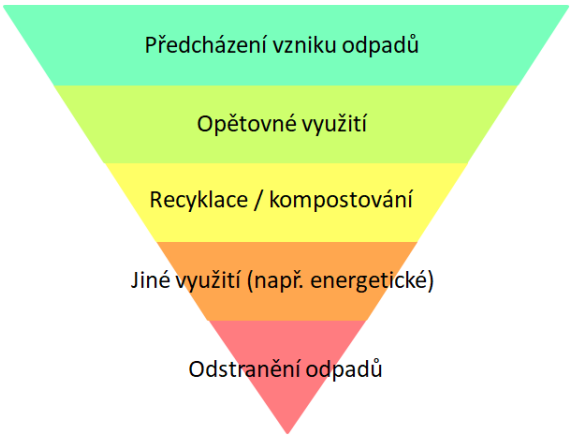
BRNO ....

podpis autora

OBSAH

1 ÚVOD

Odpadové hospodářství (OH) představuje dynamicky se rozvíjející oblast. V rozvojových zemích dochází k rapidnímu nárůstu produkce odpadu vlivem změny demografických podmínek. V rozvinutých zemích, především v západní Evropě, je trendem přechod z lineárního schématu, kdy ze vstupní suroviny vzniká produkt a odpad, k tzv. oběhovému hospodářství (nazývané rovněž cirkulární ekonomikou). Cílem je maximalizovat využití odpadu jako druhotného materiálu a minimalizovat neefektivní plýtvání hodnotného potenciálu skrytého v odpadech. Preferovaný způsob nakládání s odpady vyplývá z tzv. Hierarchie nakládání s odpady, kterou ukotvuje Zákon o odpadech [1], členění je zobrazeno na obr. 1.1. Podrobněji se této problematice věnuje [2] a naznačuje také praktické aplikace cirkulární ekonomiky.



Obr. 1.1: Hierarchie způsobů nakládání s odpady

Pro udržitelný přechod na oběhové hospodářství je třeba adekvátně přizpůsobit zpracovatelskou infrastrukturu. Ta se skládá ze sběru odpadu, dopravy, zařízení na úpravu odpadů a konečného zpracování.

Z pohledu ČR je důležitým aspektem pro plánování OH následujících let přijetí Novely zákona o odpadech č. 229/2014 [3], která stanovuje zákaz skládkování směsného komunálního odpadu (SKO) a recyklovatelných odpadů. Jedná se o významnou změnu, na kterou musejí zejména zpracovatelská zařízení včas reagovat. Úprava stávajících a případná výstavba nových zařízení bývá legislativně složitý proces, který je časově náročný. Minimální doba pro vybudování nového zařízení pro energetické využití odpadu se odhaduje na 5 až 7 let [4]. Proto je třeba v OH vycházet z kvalitních komplexních plánů, které odrážejí potřebný časový horizont. Z důvodu velkého množství neurčitých faktorů, které ovlivňují udržitelnost nových zařízení, je třeba při plánování disponovat sofistikovanými výpočtovými nástroji [5]. Nezbytnou částí

je rovněž analýza rizik [6]. Řešení případových studií v OH se neobejde bez kvalitních prognóz klíčových parametrů, mezi nejvýznamnější patří produkce a složení odpadu.

Shrnutím dříve publikovaných modelů pro odhad současné a budoucí produkce SKO se zabývala práce [7]. Je zde zmíněno 45 modelovacích přístupů a formulována čtyři základní kritéria modelování: velikost území, původ dat, nezávisle proměnné a modelovací metoda. Na základě shrnutých poznatků je navržen systém pro výběr vhodné metody pro analýzu, ve většině případů se jako nejlepší přístup jeví korelační a regresní analýza. Aplikace analýzy časových řad je doporučovaná ve speciálních případech, pokud má například podat informaci o sezónním charakteru dat. Na tento rozbor modelů navazuje [8], kde je utvořen přehled modelů užívaných pro odhad produkce SKO. Jsou uvažována totožná kritéria modelování, jako v předchozí práci. Zde mají významné zastoupení metody založené na umělé inteligenci [9], [10], přesto většina publikovaných modelů využívá regresní analýzu, kde je popisován vztah mezi socio-ekonomickými faktory a množstvím vyprodukovaného odpadu [11], [12], [13].

Bohužel při zpracování reálných dat často nejsou splněny předpoklady regresní analýzy, které formuluje např. [14] pro případ lineární regrese a [15] pro nelineární regresi. V některých případech je možné tyto podmínky nahradit podmínkami obecnějšími a sestavit zobecněný regresní model. Zobecněný model je schopen řešit např. problém heteroskedasticity v reziduích, korelované náhodné složky v modelu, jiné než normální rozdělení reziduí. Práce [16] shromažďuje zdroje informací na téma zobecněného lineárního modelu (generalized linear model - GLM) a prezentuje využití GLM v oblasti modelování ekologických systémů, jako je např. rovnováha mezi živočichy a jejich životním prostředím. Současně je diskutováno využití několika nových přístupů jako je GLM v regresních stromech a další. GLM je aplikován v článku [17], kde je navržen model a jeho následné zjednodušení pro kalibraci senzoru k dálkovému snímání. Klasický lineární model v tomto případě není vhodné použít např. kvůli silným korelacím. Zobecněné modely lze úspěšně používat i pro nelineární modely, jak ukazuje [18].

Studie [19] shrnuje možnosti využití umělých neuronových sítí jako alternativní přístup k prognózování a [20] se věnuje i dalším algoritmům umělé inteligence. Ty byly testovány z hlediska schopnosti předvídat měsíční produkci odpadu ve městě Logan, Austrálie. Na základě výsledků lze říci, že tyto algoritmy se v mnoha případech mohou stát vhodným nástrojem pro stanovení prognostických odhadů produkce odpadu.

Socio-ekonomická data, na kterých jsou regresní modely nejčastěji založeny, nejsou vždy dostupná na úrovni mikroregionů. Data zpracovávaná v případové studii (kap. 7), byla evidována právě na úrovni mikroregionů, kterým v ČR odpovídá členění území na obce s rozšířenou působností (ORP). Tento typ modelů tak pro zde zpra-

covávanou aplikaci není obecně vhodný a jeho využití je značně omezené.

V celé řadě dříve provedených analýz je jediným vysvětlovaným parametrem jednotka času, což vede na analýzu časových řad. Studie [21] porovnává modely časových řad pro prognózování množství komunálních odpadů (KO) pro následující roky, avšak tyto modely jsou aplikovatelné pouze na dostatečně dlouhé časové řady. Na podobném principu by mohly být sestaveny modely pro data na úrovni dnů při plánování provozu apod., na roční bázi není obvykle dostatečně dlouhá časová řada k dispozici. V krátkých časových řadách není možné nalézt autokorelace v datech nebo identifikovat jednotlivé složky časové řady (sezónní, cyklická, náhodná). Přesto složka trendu má často charakteristický průběh i v krátkých časových řadách.

Práce [22] se věnuje prognózování produkce SKO s ohledem na případovou studii pro rumunské město Iasi. Autoři využívají software Waste Prognostic Tool [23] a Minitab v kombinaci s regresní analýzou a analýzou časových řad. Jako nejvhodnější model trendu v datech se pro zmíněná data jeví S-křivky a to jak pro celkovou produkci SKO, tak pro jednotlivé složky SKO.

Tato práce navrhuje metodiku pro prognózování produkce odpadů na základě ročních dat dostupných pro velmi krátkou časovou řadu. Tradiční prognostické postupy pro zde zpracovávanou případovou studii selhávají a to z důvodu nedostupnosti socio-ekonomických dat na úrovni mikroregionů. Návrh postupu vychází z principů regresní analýzy s cílem modelovat trend v historických datech, kde jedinou nezávisle proměnnou je průběh v čase. Metodika navržená v této práci využívá pouze trendovou složku, kterou lze poměrně úspěšně na dostupných datech pozorovat ve všech sledovaných územních celcích (ORP, kraje, ČR). Na základě území a typu odpadu mohou data vykazovat rozdílný průběh v čase, proto je v modelu (kap. 5) uvažováno velké množství funkcí pro popis trendu. Avšak funkce ve tvaru S-křivky v tomto výběru hrají významnou roli. Problémem všech výše zmíněných přístupů je nekonzistence při agregaci dat pro územní celky a složky odpadu tak, aby součet příslušných nižších celků odpovídal celku nadřazenému. Tuto problematiku řeší [24] prostřednictvím bilancování trendu ve sledovaném roce. U takto korigovaných hodnot je zachována hierarchie územních celků a katalogových čísel odpadu. Tato práce zapadá do vývoje rozsáhlého nástroje pro prognózování Justine [25].

V následujícím textu je nejprve zaveden potřebný matematický aparát, který zahrnuje regresní analýzu, analýzu vícerozměrných dat v podobě shlukové analýzy a použité oblasti z optimalizace. Další kapitola je věnována sestavení modelu, který je následně vzhledem ke svému rozsahu a výpočetní náročnosti rozložen na dílčí problémy, které jsou řešeny postupně. V rámci případové studie je navržená metodika aplikována na data o produkci KO v České republice na úrovni ORP.

1.1 Datová základna

V této práci jsou zpracovávána data o produkci odpadu v rámci systému obce na úrovni ORP, která pochází z let 2009 - 2015 a jsou evidována s kódy nakládání A00, BN30, AN60. Kódy nakládání jsou definovány v [26]. Zdrojem dat je Informační systém odpadového hospodářství (ISOH) [27], což je rozsáhlá databáze shromažďující data o produkci a nakládání s odpady v rámci celé České republiky. Původce odpadů nebo oprávněná osoba zařazují odpady pod šestimístná katalogová čísla druhů odpadů uvedená v Katalogu odpadů [28]. Jedná se o neveřejná data, která pro Ministerstvo životního prostředí (MŽP) zabezpečuje Česká informační agentura (CENIA). Pro účely této práce, jako součást řešení projektu WtECC, datovou základnu řešitelskému pracovišti Ústavu procesního inženýrství (ÚPI) poskytlo MŽP ČR. Zpracování a prezentace neveřejných dat se řídí dohodou mezi řešitelským pracovištěm ÚPI a MŽP a může v některých bodech ovlivňovat způsob a rozsah dat prezentovaných v této práci.

Prognóza produkce odpadu je provedena pro skupiny odpadů uvedených níže:

- Odpad shromážděný separovaným sběrem
  - SKO - katalogové číslo 20 03 01
  - separovaný papír (PAP) - katalogová čísla 15 01 01, 20 01 01
  - separovaný plast (PL) - katalogová čísla 15 01 02, 20 01 39
  - separované sklo (SKL) - katalogová čísla 15 01 07, 20 01 02
  - KO* = SKO + PAP + PL + SKL[1]
- Složky SKO
  - papír v SKO (PAPsko)
  - plast v SKO (PLsko)
  - sklo v SKO (SKLsko)
  - ostatní složky SKO (OSTsko)
- Celková produkce jednotlivých typů odpadu
  - celková produkce papíru (PAPcel) = PAP + PAPsko
  - celková produkce plastu (PLcel) = PL + PLsko
  - celková produkce skla (SKLcel) = SKL + SKLsko

[1]Podle [1] je KO definován jako veškerý odpad vznikající na území obce při činnosti fyzických osob a který je uveden jako KO v Katalogu odpadů [28]. V souladu s metodikou [29] není totožné, proto je zde označeno jako KO*.

Pro ostatní katalogová čísla není prognóza produkce realizována vzhledem k cílům analýzy. Převážná produkce ostatních katalogových čísel nepochází z SKO, ale tvoří zcela nový proud z pohledu separace KO [30].

KO* bude v této práci značená část KO, která zahrnuje zde uvažované typy odpadu, tedy SKO a tříděné složky - papír, plast a sklo. Množství jednotlivých složek v SKO je určeno na základě odhadu složení SKO popsaného v [30]. Celková produkce papíru, plastu a skla zahrnuje jak vytříděnou část daného typu odpadu, tak jeho zbytkové množství v SKO.

2 REGRESNÍ ANALÝZA

Jedním z nejběžnějších přístupů aproximace závislostí veličin je regrese. Následující kapitola vznikla na základě literatury [31], [32] a [33], pokud není uvedeno jinak. Regresních modelů se využívá pro nalezení vhodného vztahu mezi vysvětlovanou (závisle) proměnnou Y a vysvětlující (nezávisle) proměnnou $\mathbf{X} = (\mathbf{x}_1, ...\mathbf{x}_n)$, kde Y je náhodná veličina.

Závislost Y na $\mathbf{X}$ vyjadřuje regresní funkce:

$$y = f(\mathbf{x}, \boldsymbol{\beta}) = E(Y \mid \mathbf{X} = \mathbf{x}), \tag{2.1}$$

kde $\mathbf{x}$ je pozorovaná hodnota vektoru $\mathbf{X}$, y je realizace náhodné veličiny Y a $\boldsymbol{\beta} = (\beta_1, ..., \beta_m)$ je vektor tzv. regresních koeficientů. Hodnota závisle proměnné y odpovídá podmíněné střední hodnotě náhodné veličiny Y.

Při vyšetřování závislosti Y na $\mathbf{X}$ na základě realizace k experimentů, získáváme $(n+1)$-rozměrný statistický soubor $((y_1, x_{1,1}, x_{1,2}, ..., x_{1,n}), ..., (y_k, x_{k,1}, x_{k,2}, ..., x_{k,n}))$ s rozsahem k. Kde y_i je pozorovaná hodnota náhodné veličiny Y_i a $\mathbf{x}_i$ vektor pozorovaných hodnot nezávisle proměnných $\mathbf{X}$, pro $i = 1, ..., k$.

Cílem regresní analýzy je tedy nalezení modelu $f(\mathbf{x}, \boldsymbol{\beta})$ na základě dat a zvoleného kritéria regrese. Parametry $\boldsymbol{\beta}$ jsou neznámé konstanty a je třeba je nahradit jejich bodovými odhady $\mathbf{b} = (b_1, ..., b_m)$ tak, aby funkce

$$\widehat{y} = f(\mathbf{x}, \mathbf{b}) \tag{2.2}$$

co nejlépe aproximovala experimentální data.

Vhodné kritérium regrese lze nalézt využitím metody maximální věrohodnosti. Za předpokladu, že chyby měření působí aditivně, lze psát:

$$y_i = f(\mathbf{x}_i, \boldsymbol{\beta}) + \varepsilon_i. \tag{2.3}$$

Pokud lze o chybách ε_i předpokládat, že jsou:

- nezávislé, tzn. i měření y_i jsou vzájemně nezávislá,
- stejně rozdělenými náhodnými veličinami,
- s konstantním rozptylem,

je možné zvolit kritérium regrese odpovídající metodě nejmenších čtverců. Metoda nejmenších čtverců je v praxi nejpoužívanější, minimalizuje se zde reziduální součet čtverců $U(\boldsymbol{\beta})$:

$$U(\boldsymbol{\beta}) = \sum_{i=1}^{k} (f(\mathbf{x}_i, \boldsymbol{\beta}) - y_i)^2. \tag{2.4}$$

Odhady $\mathbf{b}$ parametrů $\boldsymbol{\beta}$ pak minimalizují kritérium $U(\boldsymbol{\beta})$.

Odchylky:

$$\widehat{\varepsilon}_i = y_i - \widehat{y}_i \tag{2.5}$$

se nazývají rezidua.

Principem je tedy minimalizace účelové funkce, nejčastěji v podobě metody nejmenších čtverců, která vyjadřuje těsnost proložení regresní a experimentální závislosti.

Vzhledem k regresním koeficientům lze regresní funkce rozdělit na lineární a nelineární.

2.1 Lineární regrese

Lineární regresní model má následující tvar:

$$y = \sum_{j=1}^{m} \beta_j f_j(\mathbf{x}), \tag{2.6}$$

kde $f_j(\mathbf{x})$ jsou známé funkce neobsahující regresní koeficienty $\beta_1, ..., \beta_m$.

Lineární regresní model splňuje následující předpoklady:

- Pro matici realizací $\mathbf{X} = \mathbf{X}_{k,n}$ platí: $k > n$, $h(\mathbf{X}) = n$. Matice je tedy plné hodnosti.
- Náhodná veličina Y_i má střední hodnotu $E(Y_i) = \sum_{j=1}^{m} \beta_j f_{ji}$ a konstantní rozptyl $D(Y_i) = \sigma^2 > 0$ pro $i = 1, ..., k$.
- Náhodné veličiny Y_i jsou nekorelované a mají normální rozdělení pravděpodobnosti pro $i = 1, ..., k$.

Pro aproximaci dat zpracovávaných v této práci bude využita regrese nelineární, proto bude více prostoru věnováno právě nelineární regresi.

2.2 Nelineární regrese

2.2.1 Formulace nelineárního regresního modelu

Lineární regresní model je lineární kombinací modelových parametrů a platí tedy podmínka:

$$g_j = \frac{\partial f(\mathbf{x}, \boldsymbol{\beta})}{\partial \beta_j} = \text{konst}, j = 1, ..., m. \tag{2.7}$$

Pokud je alespoň pro jeden parametr β_j parciální derivace g_j jeho funkcí, jedná se o nelineární regresní model.

Některé nelineární modely, nazývané vnitřně lineární modely, lze vhodnou reparametrizací převést na lineární model. Parametry $\boldsymbol{\beta}$ jsou transformovány do nových $\boldsymbol{\gamma}$, které jsou funkčně spjaty s původními parametry $\boldsymbol{\beta}$:

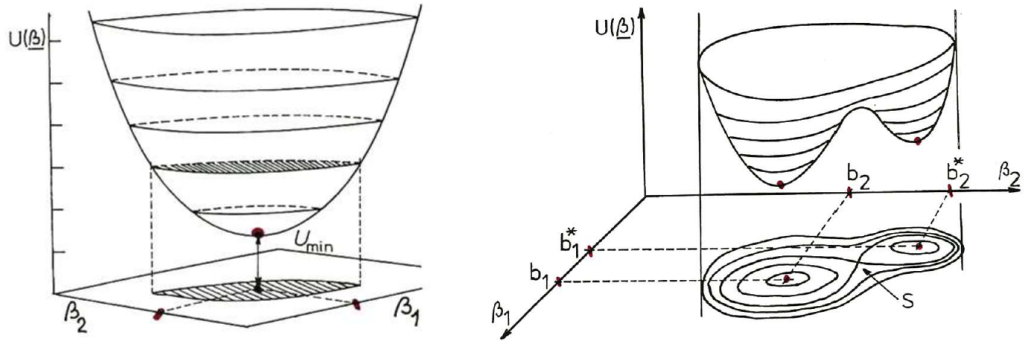
$$\boldsymbol{\gamma} = g(\boldsymbol{\beta}). \tag{2.8}$$

Pro linearizovaný model lze použít analýzu pro lineární modely.

Tvar regresní funkce se volí tak, aby co nejvíce odpovídal vyšetřované závislosti. Často se jako nelineární regresní model používá nějaká fyzikální nebo empirická závislost. Obvykle se vychází ze zkušenosti, avšak lze poměrně úspěšně využít vhodný software s databází regresních funkcí, viz kap. 6.1.1.

2.2.2 Výpočet odhadů

Pro lineární regresní modely nabývá hodnota kritéria $U(\boldsymbol{\beta})$ v prostoru odhadů svého globálního minima. Pro dva odhadované parametry β_1 a β_2 je podmínka $U(\boldsymbol{\beta})$ pro lineární model znázorněna v trojrozměrném prostoru na obr. 2.1a. U nelineárních modelů nastává komplikace v souvislosti s existencí lokálních minim a sedlových bodů. Vznikají zde složitější útvary v závislosti na nelinearitě funkce $f(\mathbf{x}, \boldsymbol{\beta})$. Obr. 2.1b znázorňuje případ funkce $U(\boldsymbol{\beta})$ se dvěma lokálními minimy $\mathbf{b}$, $\mathbf{b}^*$.



(a) Minimum účelové funkce pro lineární modely

(b) Účelová funkce nelineárního modelu se dvěma lokálními minimy

Obr. 2.1: Zobrazení minima účelové funkce [31]

Pokud je regresní model $f(\mathbf{x}, \boldsymbol{\beta})$ nelineární alespoň vzhledem k jednomu parametru β_j, vede zvolené kritérium na úlohu nelineární optimalizace, kde proměnnými jsou regresní parametry $\boldsymbol{\beta}$. Tuto úlohu lze řešit běžnými metodami pro hledání volného nebo vázaného extrému podle toho, zda jsou na regresní parametry kladena omezení.

Pro nelineární regresi mohou být užity derivační i nederivační algoritmy. Nederivační algoritmy je vhodné použít, pokud kriteriální podmínka regrese nebo regresní

funkce nejsou diferencovatelné. Pohyb směrem k extrému se pak děje pomocí řady heuristických pravidel. Derivační algoritmy patří k nejpoužívanějším, jejich nevýhodou ale je lokální konvergence, která závisí na volbě prvních odhadů $\boldsymbol{b^0}$, což může být rozhodující pro úspěch celé minimalizace. Z dobrých odhadů $\boldsymbol{b^0}$ konvergují i jednoduché metody, naopak při špatné volbě počátečního odhadu nelze nalézt minimum $U(\boldsymbol{\beta})$ žádnou z derivačních metod nebo je nalezené minimum lokálním extrémem.

2.2.3 Testování navrženého modelu

Kvalita modelu se podle [31] posuzuje s ohledem na následující kritéria:

Odhad parametrů

Kvalita nalezených odhadů se posuzuje podle intervalů spolehlivosti nebo podle rozptylů $D(b_j)$. Příčinou vysokých rozptylů parametrů je často předčasné ukončení minimalizačního procesu ještě před dosažením minima nebo značná nelinearita modelu.

Těsnost proložení

Nejznámější charakteristikou je koeficient determinace R^2:

$$R^2 = 1 - \frac{U(\mathbf{b})}{\sum_{i=1}^{k}(y_i - \overline{y})^2}, \tag{2.9}$$

kde $U(\mathbf{b})$ je reziduální součet čtverců pro bodové odhady parametrů $\mathbf{b}$, $\overline{y}$ je aritmetický průměr složek y_i. Hodnota R^2 leží v intervalu $\langle 0, 1\rangle$, vyšší hodnota poukazuje na kvalitnější regresní model.

Predikční schopnost modelu

Kritéria posouzení predikční schopnosti modelu vycházejí z rozdělení dat do dvou podskupin. Odhadovací podskupina slouží k odhadu parametrů a predikční podkupina k vyčíslení reziduálního součtu čtverců. Mezi nejjednodušší kritéria patří střední kvadratická chyba predikce:

$$MEP = \frac{1}{k}\sum_{i=1}^{k}[y_i - f(\mathbf{x}_i, \mathbf{b}_{(i)})]^2, \tag{2.10}$$

kde $\mathbf{b}_{(i)}$ je odhad parametrů určený ze všech bodů, kromě i-tého.

Kvalita dat

U nelineárních modelů slouží analýza reziduí k posouzení těsnosti proložení regresní křivky danými body. K analýze reziduí se užívá grafického zobrazení vektoru reziduí, kdy lze snadno odhalit odlehlé hodnoty v souboru reziduí, trend v reziduích, chybný model nebo vzájemnou závislost reziduí a další znaky. Numerická analýza reziduí vede ke statistickému testování. Mezi nejčastěji užívané statistiky patří střední hodnota reziduí, průměrné reziduum, směrodatná odchylka střední hodnoty reziduí, koeficient šikmosti a koeficient špičatosti.

Analýzou vlivných bodů se identifikují hodnoty, které silně ovlivňují odhadované regresní parametry a dělí se do tří základních skupin:

- hrubé chyby - obvykle vznikají v důsledku chyb při manipulaci s daty
  - odlehlé body - od ostatních dat se liší ve směru osy y, tedy závisle proměnné
  - extrémní body - liší se ve směru osy x, tedy nezávisle proměnné
  - kombinace odlehlých a extrémních hodnot
- body s vysokým vlivem - hodnoty, které byly přesně změřeny a obvykle rozšiřují predikční schopnost modelu
- zdánlivě vlivné body - vznik v důsledku nesprávně navrženého regresního modelu

Metodou pro odhalení odlehlosti pozorování v souboru s neznámým rozdělením je Dean-Dixonův test (Q-test) [34]. Hodnoty výběrového souboru jsou seřazeny vzestupně, tj. $x_1 < x_2 < ... < x_{n-1} < x_n$, a pro testování slouží testové statistiky Q_1 pokud je z odlehlosti podezřelá nejmenší hodnota a Q_n pro testování nejvyšší hodnoty:

$$Q_1 = \frac{x_2 - x_1}{x_n - x_1}, \qquad Q_n = \frac{x_n - x_{n-1}}{x_n - x_1}. \tag{2.11}$$

Testová statistika Q_1 nebo Q_n je porovnána s kritickou hodnotou Q_α. Pokud $Q_1 > Q_\alpha$ resp. $Q_n > Q_\alpha$ lze hodnotu x_1 resp. x_n označit za odlehlé pozorování na hladině významnosti α.

Vlivné body lze identifikovat užitím některých typů reziduí nebo sledováním změn, ke kterým dojde vynecháním některých bodů. Níže jsou uvedeny některé metody pro detekci vlivných bodů podle [31], [35] a [36]:

- Charekteristika DFS_{ij} vyjadřuje vliv i-tého bodu na odhad j-tého parametru:

$$DFS_{ij} = \frac{b_j - b^1_{j(i)}}{s_{(i)}\sqrt{V_{ii}}}, \tag{2.12}$$

  kde b_j je odhad parametru s indexem j, $b^1_{j(i)}$ je jednokroková aproximace odhadu s vynecháním i-tého bodu, $s^2_{(i)}$ je odhad rozptylu reziduí při vynechání

i-tého bodu a V_{ii} jsou diagonální prvky matice $\mathbf{V} = (\mathbf{J}^T\mathbf{J})^{-1}$, $J_{ij} = \frac{\partial f(x_i,\boldsymbol{\beta})}{\partial \beta_j}$. i-tý bod se považuje za vlivný, pokud je $DFS_{ij} > \frac{2}{\sqrt{k}}$.

- Jackknife reziduum $\widehat{\varepsilon}_{Ji}$ je dáno vztahem:

$$\widehat{\varepsilon}_{Ji} = \frac{\widehat{\varepsilon}_i}{s_{(i)}\sqrt{1 - P_{ii}}}, \tag{2.13}$$

kde $\widehat{\varepsilon}_i$ jsou rezidua, $s^2_{(i)}$ je odhad rozptylu reziduí při vynechání i-tého bodu, $\mathbf{P}$ je projekční matice: $\mathbf{P} = \mathbf{J}(\mathbf{J}^T\mathbf{J})\mathbf{J}^T$.

- Věrohodnostní vzdálenost LD_i je mírou vlivu i-tého bodu na odhady parametrů. Pro případ metody nejmenších čtverců má tvar:

$$LD_i = k \ln\left(\frac{U(\mathbf{b}_{(i)})}{U(\mathbf{b})}\right), \tag{2.14}$$

kde k je počet experimentů, $U(\mathbf{b})$ je reziduální součet čtverců a $U(\mathbf{b}_{(i)})$ je reziduální součet čtverců s vynecháním i-tého bodu. Je-li $LD_i > \chi^2_{1-\alpha}(2)$, je daný bod silně vlivný.

- Cookova vzdálenost CD_i byla definována v [37]. Na základě tohoto kritéria je pozorování vlivné, pokud hodnota CD_i je vyšší než 1.

$$CD_i = \frac{(\mathbf{b} - \mathbf{b}_{(i)})^T(\mathbf{J}^T\mathbf{J})(\mathbf{b} - \mathbf{b}_{(i)})}{ps^2}, \tag{2.15}$$

kde $\mathbf{b}$ je odhad regresních parametrů na základě všech dat a $\mathbf{b}_{(i)}$ je odhad parametrů s vynecháním i-tého bodu. $J_{ij} = \frac{\partial f(x_i,\boldsymbol{\beta})}{\partial \beta_j}$, p je počet hledaných parametrů v modelu a s^2 je odhad rozptylu reziduí pro model se všemi daty. Pokud je $\mathbf{b}_{(i)}$ nahrazeno lineární aproximací, Cookova vzdálenost přechází na tvar:

$$CD_i = \frac{t_i^2}{p}\frac{h_{ii}}{1 - h_{ii}}, \tag{2.16}$$

kde t_i jsou studentizovaná rezidua, h_{ii} jsou prvky matice $\mathbf{H}$, $\mathbf{H} = \mathbf{J}(\mathbf{J}^T\mathbf{J})^{-1}\mathbf{J}^T$. Jako alternativní přístup [37] navrhuje úpravu CD_i na normu změn mezi modely pro kompletní data a s vynechanými body:

$$D_i = \frac{\sum_{j=1}^{k}(\widehat{y}_j - \widehat{y}_{j(i)})^2}{ps^2}, \tag{2.17}$$

kde $\widehat{y}_j$ je odhad nezávisle proměnné v bodě $\mathbf{x}_j$, $\widehat{y}_{j(i)}$ je odhad v bodě $\mathbf{x}_j$ s vynecháním dat s indexem i. I v tomto případě orientačně platí, že je-li tento tvar Cookovy vzdálenosti $D_i > 1$, lze označit i-té pozorování jako vlivný bod. Přesto [38] doporučuje, aby byly všechny hodnoty Cookovy vzdálenosti zhodnoceny z grafického hlediska, namísto stanovení kritické hodnoty. Pokud se hodnoty Cookovy vzdálenosti pro jednotlivá data významně neliší, není třeba se jimi dále zabývat. Naopak pokud se některá hodnota odklání od ostatních, měl by být tento bod označen jako vlivný.

Správnost navrženého modelu

Pro testování správnosti navrženého modelu lze užít např. Whitův test. Testy správnosti navrženého nelineárního modelu obsahuje [39].

Souhlas s fyzikální interpretovatelností

Na odhady parametrů mohou být kladeny požadavky, které mají fyzikální smysl. Kontroluje se tedy, zda takové parametry leží v předpokládané oblasti.

3 SHLUKOVÁ ANALÝZA

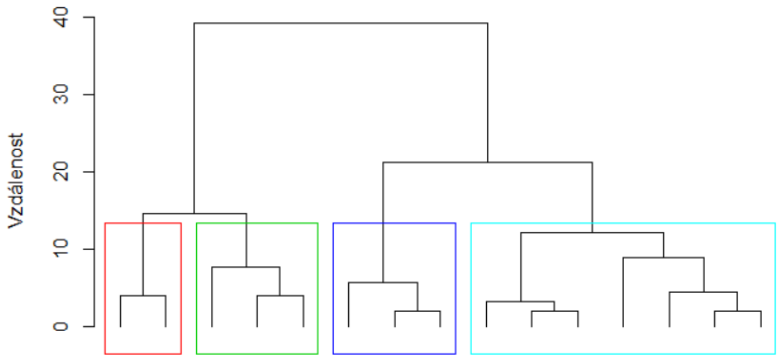
Shluková analýza je jednou z oblastí klasifikace dat, cílem shlukové analýzy je rozdělení objektů do tříd neboli shluků. V analýze shluků neexistuje trénovací množina, na základě které by byly objekty do skupin rozděleny, jako tomu může být u jiných klasifikačních metod. Pro všechny objekty je tedy neznámá příslušnost do skupin. Jedním z hlavních důvodů využití shlukové analýzy je zjednodušení dat, kdy může poskytnout zjednodušený pohled na objekty. V následujícím textu bylo čerpáno z [31] a [40].

Analýza shluků vytváří shluky podobných objektů. Meziobjektová podobnost bývá nejčastěji měřena mírou vzdálenosti, která je reprezentována metrikou. Jako příklad lze zmínit eukleidovskou a manhattanskou vzdálenost, Minkovského metriku a další. Podobnost objektů lze také měřit mírou korelace a mírou asociace.

Algoritmy shlukové analýzy se podle způsobu shlukování dělí na hierarchické a nehierarchické shlukování.

3.1 Hierarchické shlukování

Základním znakem hierarchického shlukování je hierarchické uspořádání objektů a jejich shluků. Takové shluky je možné graficky zobrazit prostřednictvím dendrogramu a konkrétní rozklad na shluky je dán vodorovným řezem. Na jedné ose jsou vyneseny jednotlivé objekty a druhá osa znázorňuje vzdálenost. Obr. 3.1 je ukázkou dendrogramu pro 15 objektů, které byly klasifikovány do 4 shluků řezem ve vzdálenosti 14. Jednotlivé shluky jsou v obrázku vyznačeny barevně.



Obr. 3.1: Ukázka dendrogramu

Princip metod hierarchického shlukování je rozdělen na aglomerační shlukování a divizní shlukování. V případě aglomeračního shlukování vytvoří dva objekty o nejmenší vzdálenosti první shluk a přepočítá se matice vzdáleností. Namísto prvků, které již

byly zařazeny, se v matici vzdáleností objevuje vzniklý shluk, jako jediný objekt. Postup se opakuje, až všechny objekty vytvoří jeden velký shluk. U divizního shlukování je postup zcela obrácený. Na počátku je množina všech objektů jediným shlukem a jeho postupným dělením se získává systém shluků a to tak dlouho, až je množina rozdělena na jednotlivé prvky.

Nejznámější metody hierarchického shlukování:

- Metoda nejbližšího souseda
  Vzdálenost mezi dvěma shluky je dána vzdáleností dvou nejbližších objektů těchto shluků. Častou nevýhodou je řetězový efekt. Spojeny mohou být shluky, které mají nejbližší objekty, ale vhledem k většině ostatních objektů nejde o nejbližší shluky.
- Metoda nejvzdálenějšího souseda
  Jde o metodou podobnou metodě nejbližšího souseda. Vzdálenost dvou shluků je tentokrát určena na základě vzdálenosti nejvzdálenějších prvků.
- Centroidní metoda
  Vzdálenost dvou shluků je dána vzdáleností jejich těžišť. Výhodou je menší ovlivnění odlehlými body.
- Wardova metoda
  Principem je minimalizace heterogenity shluků podle vnitroshlukového součtu čtverců odchylek objektů od těžiště shluků.

3.2 Nehierarchické shlukování

Většina algoritmů pro nehierarchické shlukování začíná počátečním rozkladem, který se poté upravuje přesouváním objektů mezi shluky. Nevýhodou tedy je potřeba informace o počtu shluků před procesem shlukování.

Nejčastěji se využívá algoritmus k-means. Na počátku uživatel zadává počet shluků a umisťuje centroidy pro první iteraci. Tato volba je klíčová, protože není zaručeno nalezení globálního řešení. Každý objekt je zařazen do shluku s nejmenší vzdáleností mezi objektem a centroidem shluku. Následně dojde k přepočítání centroidů těchto shluků, které se přesunou a prvky jsou přeuspořádány stejným způsobem, jako v předchozí iteraci. Postup se opakuje, dokud se poloha centroidů neustálí.

3.3 Validace shluků

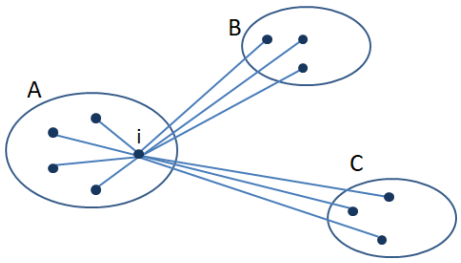
Validací shluků rozumíme metodu, která se zabývá zhodnocením vhodnosti rozdělení množiny dat na shluky. Jedním ze způsobů validace shluků jsou siluety, které poskytují informaci o kvalitě shluku [41].

Hodnota siluety $s(i)$ pro prvek i se konstruuje následovně:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}, \tag{3.1}$$

kde $a(i)$ je průměrná odlišnost prvku i od všech ostatních objektů téhož shluku. Pokud prvek i je součástí shluku A, jak ukazuje obr. 3.2, odlišnost $a(i)$ je průměrnou délkou čar spojujících prvek i s ostatními prvky shluku A. $b(i)$ je minimum z průměrných odlišností prvku i ke všem ostatním shlukům. $d(i, C)$ je průměrná vzdálenost prvku i k prvkům shluku C. Po určení $d(i, C)$ pro všechny shluky takové, že $C \neq A$, je vybrána nejmenší průměrná vzdálenost:

$$b(i) = \min_{C \neq A} d(i, C). \tag{3.2}$$



Obr. 3.2: Znázornění prvků pro výpočet siluety $s(i)$

Silueta $s(i)$ nabývá hodnot z intervalu $\langle -1; 1 \rangle$. Vyšší hodnoty $s(i)$ dosáhnou prvky, které náleží shluku o vyšším počtu prvků umístěných v menší vzdálenosti. Vysoké hodnoty $s(i)$ tedy poukazují na kompaktní a dobře oddělený shluk. Záporná hodnota naznačuje, že prvek dokonce nemusí být správně zařazen do shluku.

Siluety lze graficky znázornit tak, že na svislou osu se nanáší hodnoty $s(i)$ a na vodorovnou jsou seřazeny prvky shluku podle hodnoty $s(i)$. Pokud jsou do grafu umístěny siluety všech shluků, dostáváme siluetu rozkladu, jako je např. obr. 6.4.

4 OPTIMALIZACE

Podstatnou částí této práce je nalezení křivky, která je vhodnou aproximací dostupných dat a současně splňuje další podmínky zahrnuté v modelu 5. Nalezení nejvhodnějšího řešení problému je úlohou matematického programování (optimalizace). V následující části budou přiblíženy základní pojmy z této oblasti s využitím [42] a [43], pokud není uvedeno jinak.

Úlohu matematického programování lze formulovat následovně:

$$\min \quad f(\mathbf{x}) \tag{4.1}$$

$$\text{za podmínek} \quad g_i(\mathbf{x}) \leq 0, \quad i = 1, ..., m, \tag{4.2}$$

kde funkce $f : \mathbb{R}^n \to \mathbb{R}$ je účelová funkce, $g_i : \mathbb{R}^n \to \mathbb{R}$, $i = 1, ..., m$ jsou omezující podmínky a $\mathbf{x} = (x_1, ..., x_n)^T$ je vektor tzv. rozhodovacích proměnných. Množina $S = \{\mathbf{x} | g_i(\mathbf{x}) \leq 0, i = 1, ..., m\}$ je množinou přípustných řešení úlohy. Vektor $\mathbf{x}^*$ je označen jako optimální hodnota problému 4.1, 4.2, pokud dosahuje nejnižší hodnoty účelové funkce a splňuje všechna omezení 4.2, tedy $f(\mathbf{x}^*) \leq f(\mathbf{x}), \forall \mathbf{x} \in S$.

Speciálním případem problému 4.1, 4.2 je úloha lineárního programování, pokud je účelová funkce lineární a všechny omezující podmínky mohou být reprezentovány rovnicemi nebo nerovnicemi v lineárním tvaru. Úlohu lineárního programování lze formulovat následovně:

$$\min \quad \mathbf{c}^T\mathbf{x} \tag{4.3}$$

$$\text{za podmínky} \quad \mathbf{A}\mathbf{x} \leq \mathbf{b}, \tag{4.4}$$

$$\mathbf{x} \geq \mathbf{0}, \tag{4.5}$$

kde $\mathbf{x} = (x_1, ..., x_n)^T$ je vektor proměnných. $\mathbf{c} = (c_1, ..., c_n)^T$ a $\mathbf{b} = (b_1, ..., b_m)^T$ jsou vektory parametrů, $\mathbf{A}$ je matice typu $m \times n$ a pro prvky této matice platí $a_{ij} \in \mathbb{R}$.

4.1 Nelineární programování

Pokud účelová funkce 4.1 nebo alespoň jedna z funkcí g_i omezení 4.2 není lineární, jedná se o úlohu nelineárního programování [44], [45].

Nejjednodušší případ nastává, pokud úloha neobsahuje žádné omezení $g_i(\mathbf{x}) \leq 0$, potom se nazývá úlohou na volný extrém. V opačném případě se jedná o úlohu na vázaný extrém.

V řadě nelineárních optimalizačních úloh má účelová funkce $f(\mathbf{x})$ velký počet lokálních minim a nalezení globálního řešení tak není vždy zaručeno. Neexistuje zde univerzální metoda řešení, analytické metody často nejsou vhodné a používají se numerické metody řešení

4.1.1 Konvexní optimalizace

Konvexní programování je speciálním typem nelineární optimalizace [46]. Úloha zadaná účelovou funkcí 4.1 a omezujícími podmínkami 4.2 je úlohou konvexního programování právě tehdy, když množina přípustných řešení S je konvexní množina a účelová funkce $f(\mathbf{x})$ je konvexní funkce na S.

Konvexní množina: Množinu $S \subset \mathbb{R}^n$ nazveme konvexní množinou, jestliže pro dva libovolné body $\mathbf{x}_1, \mathbf{x}_2 \in S$ a pro libovolné $\alpha \in \langle 0; 1 \rangle$ platí:

$$\alpha \mathbf{x}_1 + (1 - \alpha)\mathbf{x}_2 \in S. \tag{4.6}$$

Konvexní množina tedy s každými dvěma svými body obsahuje i úsečku, která je spojuje.

Konvexní funkce: Nechť $f : S \to \mathbb{R}$, kde $S \subset \mathbb{R}^n$ je neprázdná konvexní množina. Funkce f je konvexní funkcí na S právě tehdy, když pro každé dva body $\mathbf{x}_1, \mathbf{x}_2 \in S$ a pro libovolné $\lambda \in \langle 0, 1 \rangle$ platí:

$$f(\lambda \mathbf{x}_1 + (1 - \lambda)\mathbf{x}_2) \leq \lambda f(\mathbf{x}_1) + (1 - \lambda) f(\mathbf{x}_2). \tag{4.7}$$

Funkce je tedy konvexní, leží-li její graf pod libovolnou sečnou. Platí-li nerovnost 4.7 jako ostrá, pak je funkce $f(\mathbf{x})$ na množině S ryze konvexní.

Předpokládejme tedy, že množina S je konvexní. Jsou-li omezení $g_i(\mathbf{x})$ v 4.2 konvexní funkce, potom jsou konvexní množiny $S_i = \{\mathbf{x} \in S | g_i(\mathbf{x}) \leq 0\}$. Průnik konvexních množin je opět konvexní množinou, proto je konvexní i množina přípustných řešení $S = \bigcap_{i=1}^{n} S_i$.

Vlastnosti úlohy konvexního programování:

- každé lokálně optimální řešení je jejím optimálním řešením,
- je-li množina optimálních řešení neprázdná, je konvexní,
- je-li $f(\mathbf{x})$ ryze konvexní, má úloha nejvýše jedno optimální řešení.

Těchto vlastností lze s výhodou využít při řešení úlohy, protože optimalizační metody často konvergují k lokálnímu řešení. V případě konvexní optimalizace není nutné rozlišovat mezi lokálními a globálními extrémy.

Příkladem konvexního programování je metoda nejmenších čtverců.

4.1.2 Některé další typy nelineární optimalizace

- Kvadratické programování:

$$\min \quad \frac{1}{2}\mathbf{x}^T\mathbf{Q}\mathbf{x} + \mathbf{c}^T\mathbf{x} \tag{4.8}$$

$$\text{za podmínky} \quad \mathbf{A}\mathbf{x} \leq \mathbf{b}, \tag{4.9}$$

  kde $\mathbf{Q}$ je pozitivně semidefinitní matice, $\mathbf{A}$ je matice parametrů a $\mathbf{c}$, $\mathbf{b}$ jsou vektory parametrů. $\mathbf{x}$ je vektor proměnných. Účelová funkce je kvadratická, omezující podmínky jsou lineární a tvoří konvexní množinu. Jedním z algoritmů, který řeší úlohu kvadratického programování, je Wolfeho algoritmus [47].
- Separabilní programování:
  Účelová funkce neobsahuje smíšené členy a funkce f je separovatelná, dá se vyjádřit ve tvaru součtu funkcí jedné proměnné:

$$\min \quad \sum_{i=1}^{n} f_i(x_i). \tag{4.10}$$

  Úlohu tohoto typu lze s dostatečnou přesností aproximovat modelem lineárního programování. Princip takové úpravy je v nahrazení každé nelineární funkce po částech lineární aproximací [48].

4.2 Celočíselná optimalizace

Obecná úloha celočíselného programování má následující tvar:

$$\min \quad f(\mathbf{x}) \tag{4.11}$$

$$\text{za podmínek} \quad g_i(\mathbf{x}) \leq 0, \quad i = 1, 2, ..., m, \tag{4.12}$$

$$x_j \in M \subseteq \mathbb{Z}, \quad j \in J, \tag{4.13}$$

kde $J \neq \emptyset$, $J \subseteq \{1, 2, ..., n\}$ a $\mathbb{Z}$ je množina celých čísel, pro x_j, $j \notin J$ platí, $x_j \in \mathbb{R}$.

Tato obecná formulace zahrnuje lineární i nelineární úlohy celočíselného programování. Linearita je dána charakterem funkcí $f, g_1, g_2, ..., g_m$. Pokud se podmínka celočíselnosti vztahuje na všechny proměnné (tzn. $J = \{1, 2, ..., n\}$), jedná se o ryze celočíselnou úlohu. V případě smíšeně celočíselné úlohy je celočíselnost požadována pouze u některých proměnných (tzn. $J \subset \{1, 2, ..., n\}$).

Speciálním typem úloh celočíselného programování je skupina bivalentního (nula-jedničkového) programování. Hodnoty mohou nabývat pouze hodnot nula nebo jedna, takže $M_j = \{0, 1\}$, $j \in J$. Taková proměnná může reprezentovat výběr z několika možností volby, zapínání a vypínání přepínačů, odpovědi typu ano/ne a spoustu dalších situací [49] nebo [50].

4.2.1 Metody řešení celočíselného programování

Metoda úplné enumerace

V případě úplně celočíselných problémů většinou existuje konečný počet přípustných řešení, proto má smysl zvažovat použití metody úplné enumerace. Princip spočívá v sestavení všech možných kombinací hodnot proměnných, vyloučení takových kombinací, které nesplňují omezení a následně výběr nejlepšího řešení na základě hodnoty účelové funkce. V mnoha praktických problémech je počet přípustných řešení příliš velký a je tedy nutné přistoupit k jinému způsobu řešení.

Metoda větví a mezí

Jedná se o iterační metodu, je založena na opakování dvou operací:

- větvení - množina přípustných řešení M (a dále její vybraná podmnožina) se rozkládá na po dvou disjunktní podmnožiny
- omezování - pro každou podmnožinu z předchozího kroku je určena dolní mez účelové funkce na této podmnožině

Pro další rozklad je zvolena podmnožina s nejnižší dolní mezí. Cílem je nalézt přípustné řešení, pro které není hodnota účelové funkce větší než dolní meze.

Metoda sečných nadrovin

Na počátku jsou dočasně zanedbány podmínky celočíselnosti, tato spojitá úloha je řešena vhodnou optimalizační metodou. Pokud optimální řešení upravené úlohy vyhovuje podmínkám celočíselnosti, jedná se také o řešení původní úlohy. V opačném případě jsou do spojité úlohy doplněna lineární omezení s těmito vlastnostmi:

- není splněno pro optimální neceločíselné řešení,
- je splněno pro přípustná řešení původního celočíselného problému.

Doplněný spojitý problém se znovu řeší a postup se opakuje.

Heuristické metody

Pokud je problém příliš složitý, není možné jej řešit exaktními metodami. Potom je nutné použít heuristické metody, které ale obecně nezaručují nalezení globálního řešení. Jako příklad takových metod lze uvést:

- Lokální hledání
- Simulované žíhání
- Genetické algoritmy

Podrobnější popis metod pro řešení úloh celočíselného programování lze nalézt např. [51].

4.2.2 Smíšené celočíselné nelineární programování

Řada praktických problémů zahrnuje jak diskrétní proměnné, tak nelinearitu v modelu. Taková úloha vede na smíšený celočíselný nelineární problém (MINLP) [52]. Dochází tak ke kombinaci obtížnosti celočíselného a nelineárního programování.

Softwary vyvinuté pro řešení MINLP využívají často dva přístupy:

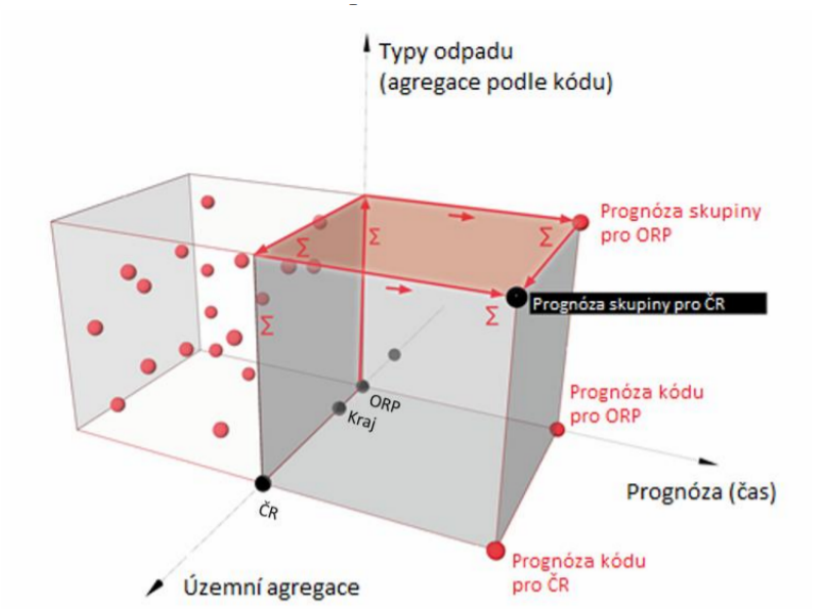
- Zobecněná Bendersova dekompozice
  Algoritmus je založen na střídavém řešení dvou úloh - smíšené celočíselné lineární úlohy a úlohy nelineárního programování.
- Metoda větví a mezí
  Metoda větví a mezí pro smíšené celočíselné lineární úlohy může být rozšířena pro případ nelineárních úloh.

Heuristické metody:
Princip heuristických metod pro řešení MINLP je často založen na zlepšování počátečního řešení $\mathbf{x}^*$. Jako $\mathbf{x}^*$ je zvoleno nějaké přípustné řešení a následně je hledáno lepší přípustné řešení.

5 MATEMATICKÝ MODEL

Tato práce se zabývá prognózou produkce odpadu, cílem je stanovit odhady budoucí produkce složek KO na úrovni ORP v ČR. Prvotní aplikace na téma prognózování řešená na ÚPI se zaměřovala na problematiku nebezpečných odpadů (NO) a byla prezentována v [24]. Na rozdíl od KO katalogová čísla spadající do NO, zpracovávaná v [24], nebyla mezi sebou provázána a prognóza byla řešena pro každé katalogové číslo zvlášť.

Data o produkci pro malé územní jednotky se často vyznačují vysokou variabilitou, což komplikuje predikci produkce, obzvláště vezmeme-li v úvahu velmi krátkou časovou řadu dostupných dat. Tuto variabilitu lze částečně potlačit agregací dat do vyšších územních celků, současně je ale třeba zachovávat vztahy mezi územními jednotkami a skupinami druhů odpadu na různých úrovních hierarchické struktury. Tedy odhad produkce vyššího územního celku odpovídá součtu produkce v uzlech, které tomuto celku náleží. Prognóza produkce v každém kraji je rovna součtu odhadů produkce v příslušných ORP a ČR je tvořena součtem krajů. Stejným způsobem funguje i vztah mezi typy odpadů, kde bývá využívána agregace některých katalogových čísel. Zmíněné vazby území a kódů odpadu by měly být zachovány pro různé provedení agregace dat. Což by bylo zaručeno pro lineární regresní model $y = a + bx$, nikoliv ale obecně pouhou analýzou trendu. Situaci graficky znázorňuje obr. 5.1, kde agregace dat je označena symbolem $\sum$ a provedení predikce dat je značeno symbolem $\rightarrow$.



Obr. 5.1: Schéma přístupu k prognózování v souvislosti s agregací dat [24]

Agregací dat se dále zabýval [53], kde byl formulován model, který vzhledem ke své náročnosti nelze implementovat a byl řešen s využitím heuristiky. Tento přístup využívá agregace dat do vyšších celků pro zlepšení kvality predikce a tvoří takové skupiny území, které vedou na nejnižší sumu kvadrátů reziduí mezi daty a odhadem.

Následující kapitola zmíněný přístup zobecňuje. Níže uvedený model připouští všechny možné agregace území, které jsou značené jako skupiny s. Model tedy uvažuje skupiny s jako kombinace území n o počtu prvků v jedné skupině $1, 2, ..., |N|$. Příslušnost území ke skupině je dána parametrem $\delta_{n,s}$. Stejným způsobem jsou definovány také všechny možné agregace katalogových čísel odpadu h, které tvoří skupiny odpadů t. Příslušnost katalogového čísla ke skupině je definována maticí $\gamma_{h,t}$.

Použité symboly

Množiny:

$i \in I$	množina let, pro které bude tvořena prognóza
$j, z \in J$	množina let s dostupnými daty, $J \subset I$
$k \in K$	množina uvažovaných tvarů regresních funkcí
$s \in S$	množina skupin vytvořených agregací území
$n \in N$	množina území
$t \in T$	množina skupin vytvořených agregací katalogových čísel
$h \in H$	množina katalogových čísel odpadů

Parametry:

$d_{j,n,h}$	data produkce odpadu s katalogovým číslem h v roce j na území n
$y_{j,s,t}$	agregovaná data o produkci pro rok j a skupiny s, t
$f_{i,s,t,k}(\mathbf{x})$	regresní funkce s indexem k pro skupiny s a t s nezávisle proměnnou i a regresními parametry danými vektorem $\mathbf{x}$
$\delta_{n,s}$	parametr indikující příslušnost území n do skupiny s
$\gamma_{h,t}$	parametr indikující příslušnost katalogových čísel odpadu h do skupin typů odpadu t
$v_{s,t}$	váha agregace do skupiny území s a skupiny katalogových čísel t
$f^*_{i,s,t,k,z}(\mathbf{x})$	regresní funkce s indexem k pro vynechaný rok z, skupiny s a t s nezávisle proměnnou i a regresními parametry danými vektorem $\mathbf{x}$
$l_{j,z}$	parametr indikující postupné vynechávání dat z
$w_{s,t,z}$	váha agregace do skupiny území s vynechaným rokem z
$Q_{s,t}$	kritická hodnota pro kritérium $D_{s,t,z}$
M	vysoká hodnota

p_k	počet regresních parametrů regresní funkce s indexem k
$E_{j,s,t}$	označení odlehlých hodnot, 0 pro odlehlé hodnoty, jinak 1
$B_{s,t,k}$	parametr pro přenesení informace o vybrané regresní funkci do druhé části modelu

Proměnné:

$m_{i,s,t}$	odhad produkce skupiny území s odpadu t v roce i
$\mathbf{x}$	vektor regresních koeficientů, dimenze vektoru $\mathbf{x}$ odpovídá nejvyššímu počtu regresních koeficientů pro jednotlivé funkce, $\dim(\mathbf{x}) = \max(p_k)$
$b_{s,t,k}$	binární proměnná pro výběr regresní funkce
$Se^2_{s,t}$	hodnota rozptylu reziduí pro skupiny s a t
$m^*_{j,s,t,z}$	odhad produkce v roce j, pro skupiny s, t s vynechaným rokem z
$D_{s,t,z}$	hodnota kritéria pro skupiny s a t a vynechaný rok z
$\varepsilon_{s,t,z}$	binární proměnná, 0 pokud je $D_{s,t,z}$ menší než kritická hodnota, jinak 1
$\alpha_{s,t}$	binární proměnná pro přepínání rostoucího a klesajícího trendu
$\alpha^*_{s,t,z}$	binární proměnná pro přepínání rostoucího a klesajícího trendu pro data s vynechaným rokem z

Návrh modelu má dvě části, které popisují kapitoly 5.1 a 5.2.

5.1 Výběr regresní funkce a identifikace odlehlých hodnot agregovaných dat

Důležitým předpokladem pro kvalitní model je nalezení vhodného tvaru regresní funkce $f_{i,s,t,k}(\mathbf{x})$, která s dostatečnou přesností popíše trend historických dat, s nezávisle proměnnou i a regresními koeficienty $\mathbf{x}$. Zde je třeba brát v úvahu fakt, že data mohou vykazovat zcela rozdílný trend na různých územích dokonce i v rámci jednoho druhu odpadu. Za tímto účelem v modelu figuruje binární proměnná $b_{s,t,k}$, která zapíná a vypíná tvary funkcí podle aktuálních dat. Zachovává se tedy možnost volby regresní funkce z množiny funkcí K pro různé časové řady.

Součástí této části modelu je i identifikace vlivných bodů, které zkreslují regresní model s využitím Cookovy vzdálenosti [37].

Účelová funkce

Účelová funkce 5.1 minimalizuje vážený součet kvadrátů chyb a sestává ze dvou základních částí. V první z nich je tvořen odhad na základě všech dostupných dat,

tuto část násobí dostatečně vysoká hodnota, aby byla minimalizována s větší prioritou. Toto opatření zajistí, že tvar regresní funkce je zvolen na základě kompletní datové sady. Druhá část minimalizuje rozdíl dat a modelu, tentokrát ale pro data s postupným vynecháváním vždy jednoho bodu prostřednictvím vektoru $l_{j,z}$. Vynechaný bod je označen indexem z, tzn. například pro vynechání prvního bodu má vektor tvar $l_{j,1} = \underbrace{(0, 1, 1, 1, ...)}_{|J|}$. Váhy $v_{s,t}$ a $w_{s,t,z}$ zajišťují znormování dat, aby bilance probíhala rovnocenně pro každou časovou řadu. Veškeré vstupní údaje na různých stupních hierarchie se díky systému vah stávají stejně významné. Součet kvadrátů chyb v účelové funkci probíhá přes množinu J, tedy roky s dostupnými daty, model je ale tvořen pro všechny prvky množiny I.

$$\begin{aligned} \min \quad & M \sum_{s\in S}\sum_{t\in T}\sum_{j\in J} v_{s,t}(m_{j,s,t} - y_{j,s,t})^2 + \\ & \sum_{s\in S}\sum_{t\in T}\sum_{j\in J}\sum_{z\in J} l_{j,z} w_{s,t,z}(m_{j,s,t,z} - y_{j,s,t})^2. \end{aligned} \tag{5.1}$$

Omezující podmínky

Podmínka 5.2 definuje data $y_{j,s,t}$ pro skupiny s a t v roce j. $\delta_{n,s}$ je matice udávající příslušnost území n do skupiny s a $\gamma_{h,t}$ je matice příslušnosti katalogového čísla h do skupiny odpadů t. Zmíněné matice mají takový tvar, že definují všechny možné agregace.

$$y_{j,s,t} = \sum_{h\in H}(\sum_{n\in N} d_{j,n,h}\delta_{n,s})\gamma_{h,t}, \forall j \in J, \forall s \in S, \forall t \in T. \tag{5.2}$$

Rovnice 5.3 definuje model produkce $m_{i,s,t}$ pro rok i a skupiny s a t prostřednictvím regresních funkcí $f_{i,s,t,k}(\mathbf{x})$. Index k značí jednotlivé regresní funkce. Proměnná $b_{s,t,k}$ nabývá hodnot 0 nebo 1 a vzhledem k podmínce 5.4 je pro každou skupinu s a t aktivní vždy pouze jediná funkce f, která nejlépe charakterizuje daná data.

$$m_{i,s,t} = \sum_{k\in K} f_{i,s,t,k}(\mathbf{x}) b_{s,t,k}, \forall i \in I, \forall s \in S, \forall t \in T, \tag{5.3}$$

$$\sum_{k\in K} b_{s,t,k} = 1, \forall s \in S, \forall t \in T, \tag{5.4}$$

$$b_{s,t,k} \in \{0, 1\}, \forall s \in S, \forall k \in K, \forall t \in T. \tag{5.5}$$

5.6 omezuje odhad produkce $m_{i,s,t}$ minimálním $u_{s,t}$ a maximálním $U_{s,t}$ množstvím. Tyto meze zabrání nereálnému poklesu nebo nárůstu odhadu produkce ve sledovaném období I.

$$u_{s,t} \leq m_{i,s,t} \leq U_{s,t}, \forall i \in I, \forall s \in S, \forall t \in T. \tag{5.6}$$

Funkce z množiny K nemusí mít obecně monotónní průběh, ale vzhledem ke zpracovávaným datům je tato vlastnost vyžadována. Monotónní trend zajistí podmínka 5.7, $\alpha_{s,t}$ je binární proměnná, nabývá hodnoty 0 pro rostoucí trend a 1 pro klesající. Vzhledem k tomu, že podmínka připouští i rovnost nule, model může mít i konstantní trend nebo jeho část.

$$(2\alpha_{s,t} - 1)(m_{i+1,s,t} - m_{i,s,t}) \leq 0, \forall s \in S, \forall t \in T, \forall i = 1, .., |I| - 1, \tag{5.7}$$

$$\alpha_{s,t} \in \{0, 1\}, \forall s \in S, \forall t \in T. \tag{5.8}$$

Pro identifikaci vlivných bodů je jako kritérium zvolena Cookova vzdálenost 2.17, která je založena na sledování změn modelů při vypouštění jednotlivých bodů z z datové sady. Součet kvadrátů rozdílů mezi modelem určeným ze všech bodů a po vynechání je normován rozptylem reziduí $Se^2_{s,t}$ a počtem parametrů regresní funkce p_k, $b_{s,t,k}$ určí počet parametrů pouze pro vybranou regresní funkci z části 5.3.

$$D_{s,t,z} = \frac{\sum_{j \in J}(m_{j,s,t} - m^*_{j,s,t,z})^2}{Se^2_{s,t} \sum_{k \in K} p_k b_{s,t,k}}, \forall s \in S, \forall t \in T, \forall z \in J. \tag{5.9}$$

5.10 - 5.14 zajišťují výpočet hodnot nutných pro určení Cookovy vzdálenosti v bodech z. Omezení 5.10 tvoří regresní modely $m_{j,s,t,z}$ s vynecháním roku z a to se stejným tvarem regresní funkce, jako byla zvolena pro kompletní data v podmínce 5.3. To, že tvar účelové funkce je vybrán pro kompletní data, je zajištěno hodnotou M v účelové funkci 5.1. 5.11 omezuje odhad produkce minimální a maximální hodnotou, 5.12 a 5.13 opět zachová monotónní průběh trendu.

$$m^*_{j,s,t,z} = \sum_{k \in K} f^*_{j,s,t,k,z}(\mathbf{x}) b_{s,t,k}, \forall j \in J, \forall s \in S, t \in T, \forall z \in J, \tag{5.10}$$

$$u_{s,t} \leq m_{i,s,t,z} \leq U_{s,t}, \forall i \in I, \forall s \in S, \forall t \in T, \tag{5.11}$$

$$(2\alpha^*_{s,t,z} - 1)(m_{i+1,s,t,z} - m^*_{i,s,t,z}) \leq 0, \forall s \in S, \forall t \in T, \forall z \in Z, \forall i = 1, .., |I| - 1, \tag{5.12}$$

$$\alpha^*_{s,t,z} \in \{0, 1\}, \forall s \in S, \forall t \in T, \forall z \in J. \tag{5.13}$$

5.14 vypočte rozptyl reziduí pro model se všemi dostupnými daty.

$$Se^2_{s,t} = \frac{1}{|J|} \sum_{j \in J} \left((m_{j,s,t} - y_{j,s,t}) - \frac{\sum_{j \in J}(m_{j,s,t} - y_{j,s,t})}{|J|} \right)^2, \forall s \in S, \forall t \in T. \tag{5.14}$$

Následující omezení 5.15 - 5.17 porovnávají hodnotu Cookovy vzdálenosti $D_{s,t,z}$ s kritickou hodnotou Q. Pokud je Cookova vzdálenost větší, než kritická hodnota, je binární proměnná $\varepsilon_{s,t,z}$ rovna 1, jinak 0. Pokud je $\varepsilon_{s,tz} = 1$, je pro skupiny s a t bod s indexem z označen jako vlivný.

$$D_{s,t,z} \leq Q_{s,t} + M\varepsilon_{s,t,z}, \forall s \in S, \forall t \in T, \forall z \in J, \tag{5.15}$$

$$D_{s,t,z} > Q_{s,t}\varepsilon_{s,t,z}, \forall s \in S, \forall t \in T, \forall z \in J, \forall t \in T, \tag{5.16}$$

$$\varepsilon_{s,t,z} \in \{0, 1\}, \forall s \in S, \forall z \in J, \forall t \in T. \tag{5.17}$$

Pro následující krok popsaný v kap. 5.2 jsou do parametrů, $B_{s,t,k}$, $E_{j,s,t}$ uloženy proměnné z první části modelu. Přenáší se tak informace o vybrané regresní funkci a o odlehlých hodnotách, které budou pro finální výpočet modelu vynechány.

$$B_{s,t,k} = b_{s,t,k} \tag{5.18}$$

$$E_{j,s,t} = (1 - \varepsilon_{s,t,z}) \tag{5.19}$$

5.2 Výpočet trendu v datech a závěrečná bilance

Model vypočte hodnoty regresních funkcí, jejichž podoba už je známa z předchozího kroku 5.1, včetně případného odstranění odlehlých hodnot. Navíc probíhá bilance odhadu produkce do vyšších územních celků a skupin katalogových čísel odpadů, viz obr. 5.1. Součet odhadů celků z níže položených skupin hierarchické struktury tak odpovídá odhadu ve vyšším celku a tím je zajištěn soulad mezi agregovanými daty.

Účelová funkce

Minimalizován je opět vážený součet kvadrátů chyb 5.20. Data $y_{j,s,t}$ pro skupiny s a t jsou určena agregací evidovaných dat již z předchozí části. Parametr $E_{j,s,t}$ zajistí vypuštění odlehlých pozorování, u těchto bodů nabývá hodnoty 0 a odklon modelu od vstupních dat se tedy v účelové funkci nepromítne. Data jsou i v této části normována váhou $v_{s,t}$.

$$\min \sum_{s \in S} \sum_{t \in T} \sum_{j \in J} E_{j,s,t} v_{s,t} (m_{j,s,t} - y_{j,s,t})^2. \tag{5.20}$$

Omezující podmínky

Regresní model je zde určen s regresními funkcemi zvolenými již v rámci první části modelu a prostřednictvím parametru $B_{s,t,k}$ se tato informace přenáší do podmínky 5.21. Odhad produkce se i v tomto případě bude pohybovat v rámci mezí $u_{s,t}$ a $U_{s,t}$, 5.22.

$$m_{i,s,t} = \sum_{k \in K} f_{i,s,t,k}(\mathbf{x}) B_{s,t,k}, \forall i \in I, \forall s \in S, t \in T, \tag{5.21}$$

$$u_{s,t} \leq m_{i,s,t,z} \leq U_{s,t}, \forall i \in I, \forall s \in S, \forall t \in T. \tag{5.22}$$

Každý nižší územní celek je součástí vyššího územního celku a to platí i pro skupiny katalogových čísel odpadu. Součet takových nižších jednotek příslušících do jisté

skupiny odpovídá odhadu vyšší skupiny, které náleží. Tato bilance v konkrétním roce $i_0 \in I$ je zajištěna následujícími dvěma podmínkami 5.23 a 5.24.

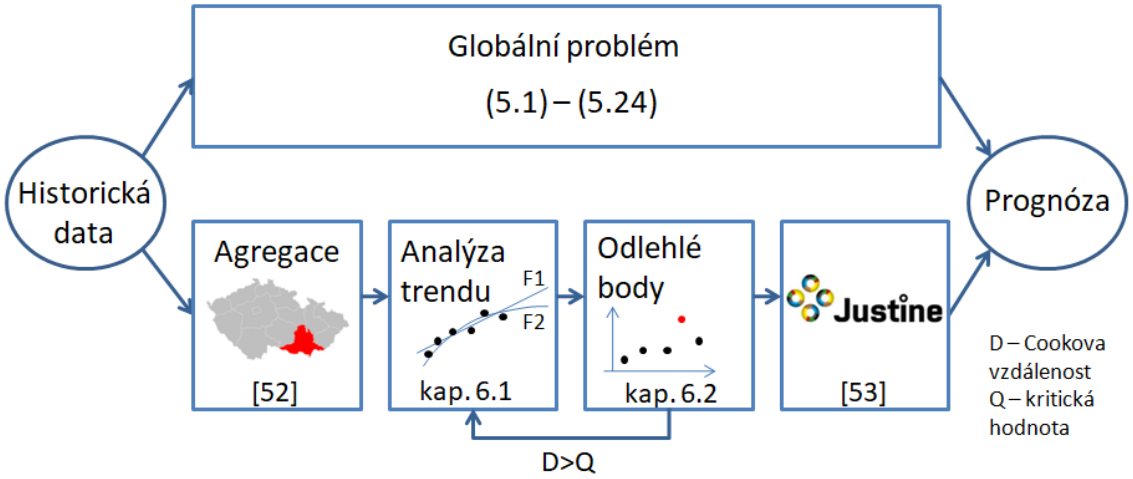
$$m_{i_0,s,t} = \sum_{n \in N} m_{i_0,n,t} \delta_{n,s}, \forall s \in S, \forall t \in T, \tag{5.23}$$

$$m_{i_0,s,t} = \sum_{h \in H} m_{i_0,s,h} \gamma_{h,t}, \forall s \in S, \forall t \in T. \tag{5.24}$$

Návrh matematického modelu tak, jak byl v této kapitole uveden, představuje úlohu smíšeného celočíselného nelineárního programování. Vzhledem k rozsahu a náročnosti této úlohy, která čítá statisíce reálných či binárních proměnných, je v následující kapitole 6 navržen systém dekompozice modelu a navazuje ukázkou aplikace tohoto přístupu na datech KO v kapitole 7.

6 DEKOMPOZICE GLOBÁLNÍ ÚLOHY

Matematický model, tak jak byl uveden v kapitole 5, přestavuje komplikovaný problém, který je pro potřeby zde zpracovávané problematiky rozložen na dílčí úlohy zajišťující řešitelnost v reálném čase. Namísto jednoho rozsáhlého globálního problému se řeší problematika agregace dat, analýzy trendu a detekce vlivných bodů, což utvoří vstup pro nástroj Justine [54], který je na ÚPI dlouhodobě vyvíjen. Tento přístup graficky znázorňuje obr. 6.1. Data mohou být vstupem přímo pro "globální problém" nebo projít postupně jednotlivé části. Pokud je při testování vlivných bodů identifikován odlehlý bod tzn. Cookova vzdálenost D překročila kritickou hodnotu Q, data se vrací k analýze trendu pro výpočet nového modelu bez vadných hodnot.



Obr. 6.1: Schematické znázornění dekompozice globální úlohy

Otázka agregace byla v zájmu kolektivu autorů práce [53]. Zde bylo navrženo sčítání dat do takových skupin, které vytvoří co nejkvalitnější data ve smyslu nejmenších kvadrátů chyb pro navržený model. Počet katalogových čísel odpadů a území náležící jedné skupině zde byl omezen za podmínky, aby se každé dvě skupiny od sebe dostatečně lišily. Tím bylo dosaženo výrazného snížení počtu skupin agregovaných dat oproti datům definovaným v modelu 5.2.

Tato práce se bude dále podrobněji věnovat problematice analýzy trendu a to zejména z hlediska výběru vhodné regresní funkce, viz kap. 6.1.1. Další části, označené jako "Odlehlé body", se věnuje kap. 6.2. Tato problematika je velmi důležitá z hlediska zachování kvality dat. Zde bude vybráno kritérium pro detekci vlivných bodů u nelineární regrese. V závěru odhady hodnot produkce vstupují do nástroje Justine, který vznikl v rámci disertační práce [55].

Cílem této práce je tedy popsat globální problém a jeho následné rozložení na podúlohy. Práce se pak v detailu zabývá vybranými podúlohami (Analýza trendu, Odlehlé body). Zbývající části jsou převzaty z již ukončených prací.

6.1 Analýza trendu

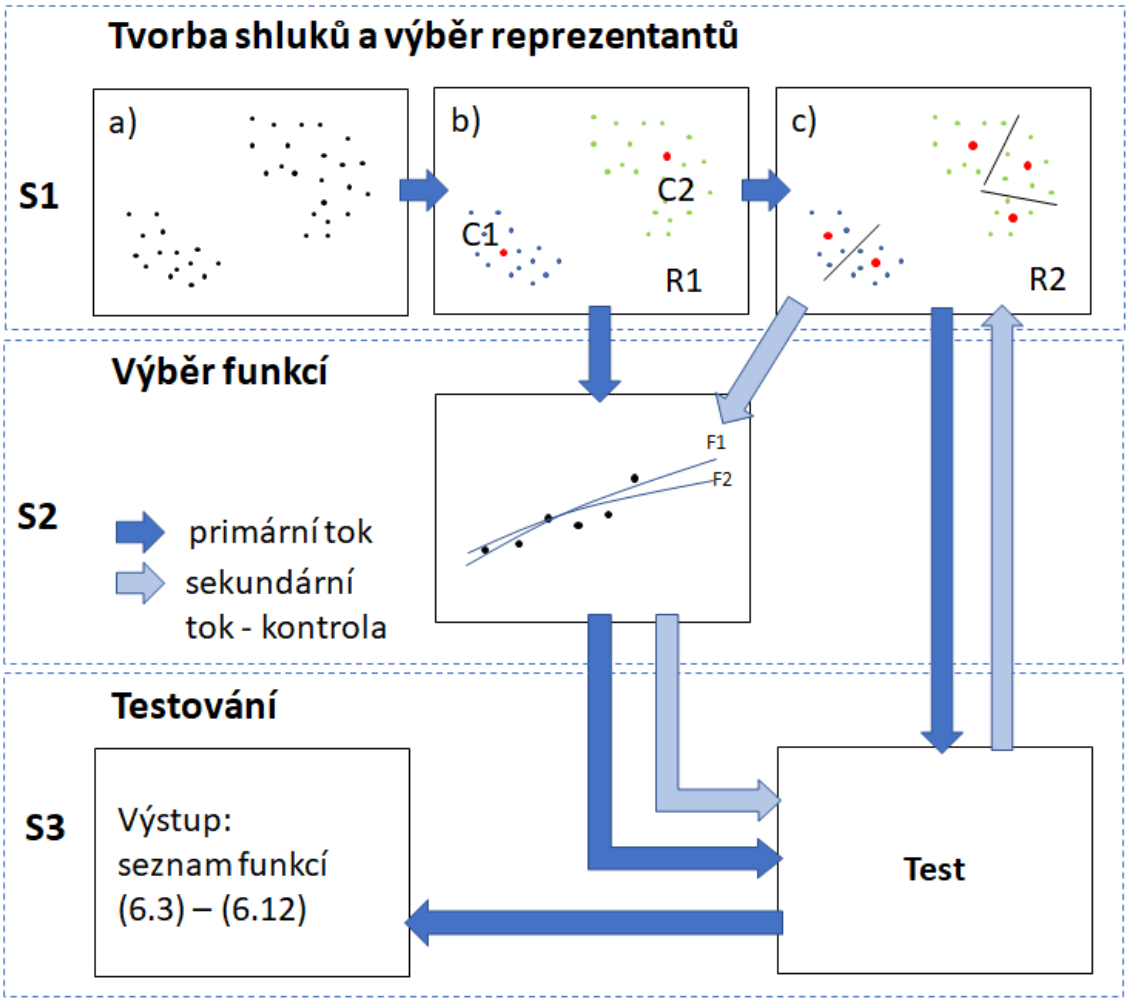
6.1.1 Volba regresní funkce

Cílem následující části je návrh vhodných regresních funkcí, které budou popisovat trend v historických datech, a tím významně snížit množství funkcí $f_{i,s,t,k}(\mathbf{x})$ z množiny K, které budou vstupovat do výpočtu. V této části je tedy řešena část 5.3 - 5.5 z modelu 5.1. Představený výpočtový systém bude testován na 12 skupinách odpadů (viz 1.1) pro 206 mikroregionů České republiky, což odpovídá 2472 časovým řadám, pro které má být nalezen regresní model. Tuto poměrně časově náročnou analýzu lze zefektivnit nalezením podobností mezi producenty a zmenšit problém pouze na podmnožinu producentů [56]. Navržený přístup je aplikován na data pouze z let 2009 - 2014. Poslední bod časové řady, rok 2015, je zachován pro závěrečné zhodnocení kvality modelů.

Metodika je založena na 3 základních krocích, následující popis doprovází schematické znázornění na obr. 6.2:

1. Shluková analýza (S1)
   Shlukovací přístup je rozdělen do několika bodů:
   a) Příprava dat - diference v časové řadě.
   b) Vytvoření shluků ze všech vstupních dat a výběr jejich reprezentantů (R1). Tito reprezentanti vstupují do kroku S2.
   c) Vytvoření podshluků ze shluků fáze S1b) a nalezení jejich reprezentantů (R2). Tito reprezentanti vstupují do kroku S3.
2. Výběr trendových funkcí (S2)
3. Testování výběru z kroku S2 na reprezentantech R2 (S3)

Na základě reprezentantů R1 je utvořen výběr regresních funkcí, které jsou vhodné právě pro tato data. Následující testování probíhá na reprezentantech podshluků R2 s využitím již vybraných funkcí pro R1, v obr. 6.2 je tato část označena jako "primární tok". Smyslem je ověřit, zda pro každý prvek shluku lze nalézt vhodnou regresní funkci z utvořeného seznamu funkcí pro daný shluk. Pokud vybrané funkce nepopisují testovaná data kvalitně, postup se vrací ke kroku S1c), tentokrát ale s reprezentanty R2, což je znázorněno jako "sekundární tok - kontrola".



Obr. 6.2: Schematické znázornění postupu pro výběr regresní funkce [56]

Shluková analýza (S1)

Důvodem tohoto kroku je nutnost snížení počtu scénářů, pro které bude hledán tvar regresní funkce. Za tímto účelem byla využita shluková analýza, přičemž se shlukují časové řady ze všech ORP a typů odpadu na základě podobného průběhu v čase. V prvním kroku byla každá řada historických dat normována Čebyševovou normou na interval $\langle 0;1\rangle$ podle předpisu:

$$\tilde{y}_{j,s,t} = \frac{|y_{j,s,t}|}{\max_j |y_{j,s,t}|}, \quad \forall j \in J, \forall s \in S, \forall t \in T, \tag{6.1}$$

kde $y_{j,s,t}$ jsou data produkce odpadu. Cílem této úpravy je znormování dat tak, aby nezáleželo na velikosti produkce, seskupeny tak budou časové řady s podobným průběhem. To umožní následně nové datové řady zařadit na základě jejich charakteru do již vytvořených a otestovaných shluků funkcí.

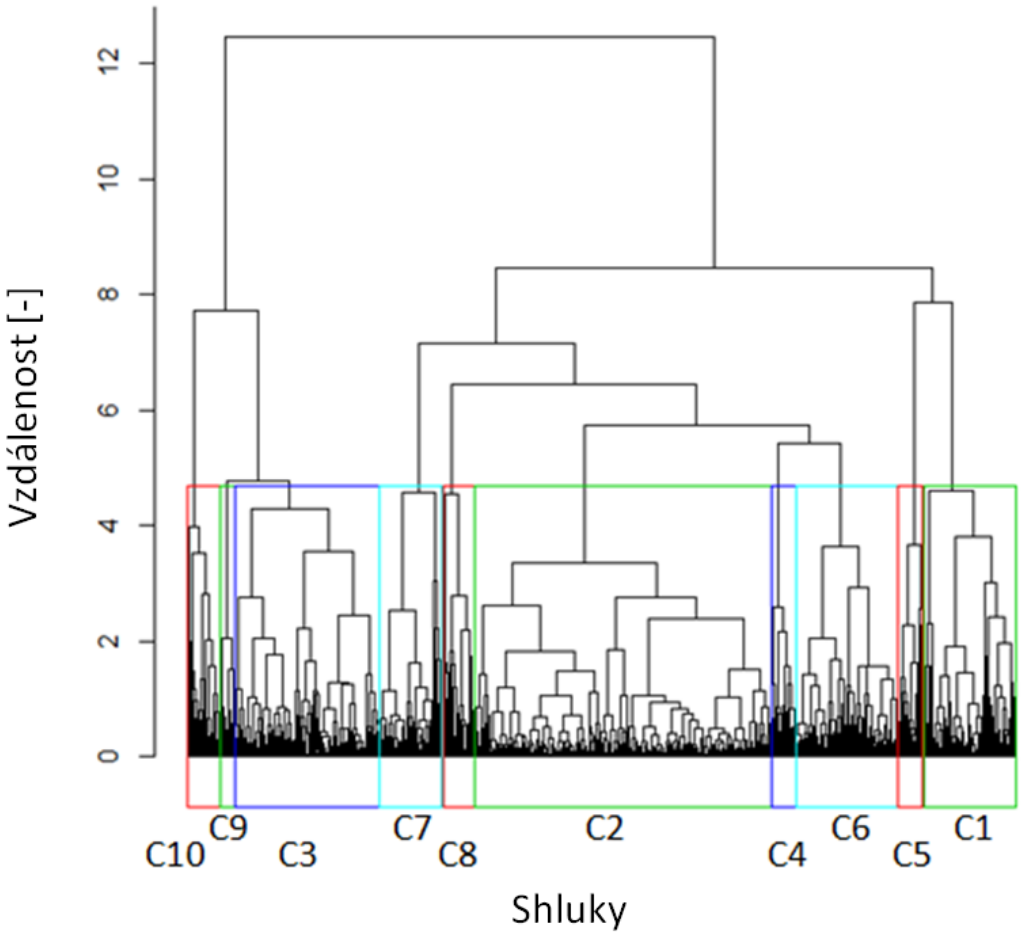
Proměnnou vstupující do shlukové analýzy jsou diference z normovaných dat:

$$\tilde{\Delta}_{j,s,t} = \tilde{y}_{j+1,s,t} - \tilde{y}_{j,s,t}, \quad \forall j \in J, \forall s \in S, \forall t \in T. \tag{6.2}$$

Tyto diference jsou využity pro sestavení matice vzdáleností mezi časovými řadami. Aktuálně časová řada disponuje sedmi body, viz kap. 1.1. S výhledem na další aplikace této metody, kdy budou data přibývat, byla zvolena d_1 metrika (Manhattanská). Ve vyšších dimenzích se běžně používaná euklidovská metrika chová jinak a nemá schopnost zde popsat rozdíly v datech. Tento jev se může projevit tak, že se rozdíl vzdáleností nejbližšího a nejvzdálenějšího souseda blíží nule. Na základě [57] je vhodnější použít právě d_1 metriku.

Shluková analýza nabízí řadu metod, které jsou v základním dělení řazeny k hierarchickým a nehierarchickým (viz kap. 3). Vhodnost možných metod byla experimentálně ověřena a na základě hodnot siluet (viz níže) je uspokojivou volbou Wardova metoda jako zástupce hierarchických metod, z nehierarchických potom k-means.

O využití Manhattanské metriky pro Wardovu metodu pojednává [40]. Výsledný dendrogram se zvýrazněním jednotlivých shluků je zobrazen na obr. 6.3. Pro určení vhodného počtu shluků není k dispozici univerzální metodika. Obecně se doporučuje vést řez dendrogramu v místě, které vede na výrazně větší vzdálenost prvků ve shluku, oproti řezu, který by odpovídal většímu počtu shluků. V následujícím dendrogramu 6.3 je zvolen řez tvořící 10 shluků – C1 až C10.



Obr. 6.3: Dendrogram rozkladu na shluky

Zařazení jednotlivých typů odpadu do shluků je shrnuto v tab. 6.1. Procentuální údaj vyjadřuje procento časových řad pro daný typ odpadů, které jsou přiřazeny do jednotlivých shluků. Protože zde uvažované kategorie odpadů se přímo olivňují, nelze v této tabulce pozorovat, že by se některé typy odpadu separovaly do zvláštních shluků. Jako tomu bylo v případě výsledků v [56], kde bylo možné shledat např. podobný charakter v produkci papíru a plastu, naopak bioodpad se zcela lišil od ostatních typů odpadu a vytvářel vlastní shluky.

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	$\sum_{C_i}$
KO	8,25	54,37	19,42	0,97	1,94	5,83	4,37	1,94	0,97	1,94	100
SKO	6,31	51,94	20,87	0,97	1,94	5,83	7,28	1,94	0,97	1,94	100
PAP	21,36	9,22	8,74	14,08	6,31	25,24	5,83	3,40	0,97	4,85	100
PL	6,80	15,05	7,77	3,40	3,40	54,37	1,94	2,91	0,00	4,37	100
SKL	17,48	27,18	7,77	6,31	3,88	22,33	7,77	4,37	0,00	2,91	100
PAPsko	11,65	16,02	22,82	0,97	4,85	1,94	21,36	7,28	7,77	5,34	100
Plsko	14,56	34,47	24,27	1,94	2,43	1,94	10,19	4,37	1,46	4,37	100
SKLsko	12,14	18,93	26,21	2,43	3,88	3,88	9,71	8,74	4,85	9,22	100
OSTsko	7,77	54,37	17,96	0,97	1,94	5,34	6,31	1,46	0,97	2,91	100
PAPcel	10,19	45,63	16,02	1,94	2,43	8,25	6,31	4,37	1,46	3,40	100
Plcel	8,25	52,91	17,96	0,49	1,46	7,28	4,37	2,43	0,97	3,88	100
SKLcel	9,71	50,49	19,42	0,97	1,94	6,31	5,83	2,43	0,97	1,94	100

Tab. 6.1: Příslušnost jednotlivých typů odpadů do shluků [%]

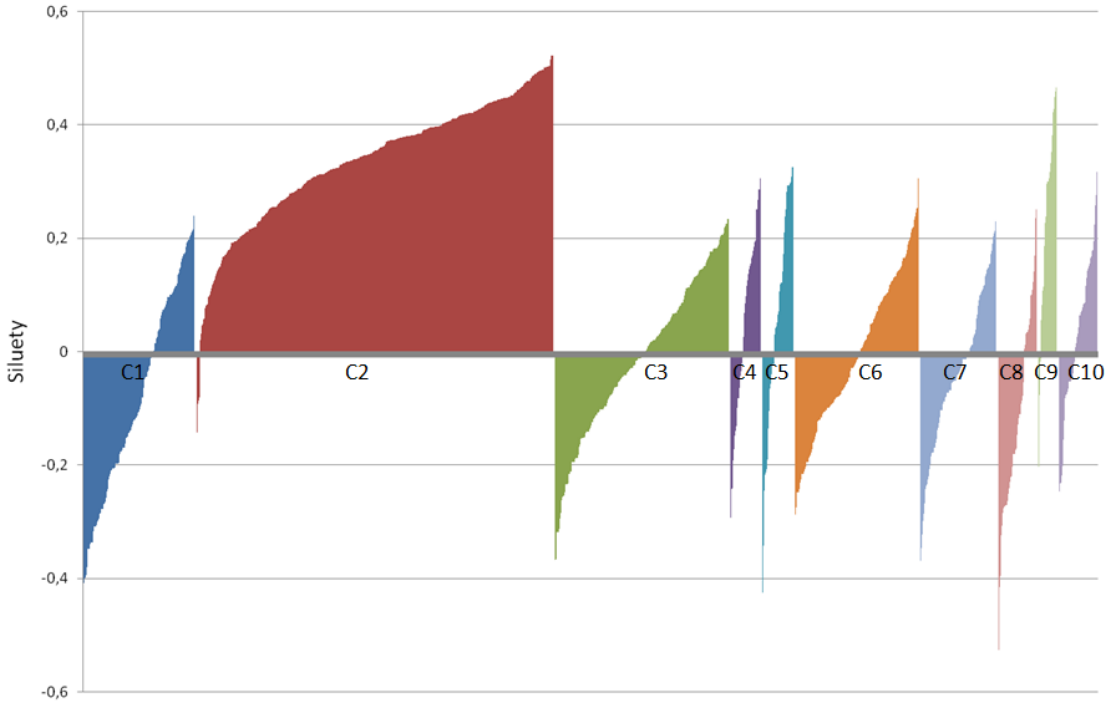
Následuje výběr reprezentantů za účelem snížení počtu časových řad, které mají být testovány pro nalezení vhodného modelu a budou tak v testování zastupovat svůj shluk. Graficky je tato situace znázorněna na obr. 6.2 v části 1b) se zvýrazněným prvkem, který byl zvolen jako reprezentant příslušného shluku (R1). Reprezentanta každého shluku lze nalézt jako prvek, který má od všech ostatních nejmenší vzdálenost.

Regresní modely jsou hledány právě pro reprezentanty shluků R1. Nicméně model by měl být zvolen tak, aby byl vhodný pro všechny časové řady, které tomuto shluku náleží. Za účelem pokrytí celé množiny dat a variability shluku jsou vybrané modely testovány a posouzeny na zvolených reprezentantech jednotlivých podshluků R2, viz obr. 6.2 1c). Tzn. každý základní shluk, na obr. 6.2 rozlišeny barevně, byl dále rozdělen, čímž získáváme více reprezentantů pro závěrečné zhodnocení.

Počet zmíněných podshluků (obr. 6.2, S1c)) a tím i počet testovaných reprezentantů R2 určuje kompaktnost každého shluku. Jinak řečeno, testování kvalitnějšího shluku probíhá pro nižší počet reprezentantů R2. Pro méně kompaktní shluk je nutné vybrat více reprezentantů R2. Pro zhodnocení kvality shluků jsou využity siluety (obr. 6.4).

Silueta $s_{k,j}$, kde j je prvek (časová řada) shluku k, nabývá hodnot z intervalu $\langle -1; 1 \rangle$ pro každý prvek j. Vyšší hodnoty $s_{k,j}$ dosáhnou prvky, které náleží shluku o vyšším počtu prvků umístěných v menší vzdálenosti. Vysoké hodnoty $s_{k,j}$ tedy poukazují na kompaktní a dobře oddělený shluk. Záporná hodnota naznačuje, že prvek dokonce nemusí být správně zařazen do shluku. Z následujícího grafického

výstupu je zřejmé, že v případě shluků C2 a C9 se jedná o dobře zařazená data a pro testování zde bude vybráno méně reprezentantů R2, než oproti např. shluku C1.



Obr. 6.4: Siluety rozkladu na shluky

Pro shluky označené jako C1 – C10 byl tedy stanoven počet reprezentantů v rozmezí 3 až 5 na základě průměrné hodnoty siluet ze všech prvků shluku, a to následujícím způsobem, viz tab. 6.2.

Počet reprezentantů R2	Průměrná hodnota siluet s_{kj} [-]
5	<0
4	$\langle$ 0; 0,2 $\rangle$
3	>0,2

Tab. 6.2: Počet podshluků na základě hodnot siluet shluků

Jednotlivým shlukům odpovídají následující počty reprezentantů R2 (tab. 6.3).

Shluk	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
Průměrná hodnota s_{kj} [-]	-0,08	0,34	-0,01	0,05	0,06	-0,01	-0,04	-0,12	0,24	0,03
Počet reprezentantů R2	5	3	5	4	4	5	5	5	3	4

Tab. 6.3: Počet podshluků na základě hodnot siluet shluků

Výběr funkcí (S2)

Každý z 10 shluků zastupuje jeden reprezentant (R1) na základě nejmenší součtové vzdálenosti od všech ostatních prvků shluku, viz obr. 6.2 S1b). Pro těchto 10 časových řad (data 2009–2014) je hledána trendová (regresní) funkce s využitím softwaru TableCurve 2D [58]. Tento software obsahuje databázi více než 3600 regresních funkcí a velmi rychle navrhne vhodné modely pro vložená data. Bohužel v jednom výpočetním běhu lze podrobit analýze pouze jednu datovou sadu a právě z tohoto důvodu je nutné provést shlukování dat. Databáze regresních funkcí v programu TableCurve 2D jsou rozděleny do 8 kategorií: Linear, Simple, Robust Straight Line, Poly/Ratl, WaveForm, Peak, Transition, Kinetic, přičemž některé funkce mohou být zařazeny ve více skupinách (např. Linear a Simple). Vzhledem k charakteru modelů jsou uvažovány pouze funkce z kategorií Linear, Simple, Peak, Transition equations. Důvodem je požadavek na monotónní trend v případě potřeby delšího prognózovaného období. V testu je sledován přijatelný časový horizont (okolo 10 let od posledních dostupných dat) s ohledem na prognózu produkce odpadu. Připuštěny tedy jsou i modely, které porušují podmínku monotonosti mimo zmíněný časový interval.

Pro každý shluk je na základě reprezentantů R1 zvoleno 12 různých tvarů regresních funkcí (3 nejlepší modely pro každou uvažovanou skupinu funkcí). Výběr modelů tvoří různé tvary funkcí s nejvyšším možným koeficientem determinace. Ve většině případů dosahuje koeficient determinace hodnot alespoň 0,8 u dvou reprezentantů se nejvyšší koeficient determinace pohybuje okolo 0,5. Výrazně horší hodnota koeficientu determinace náleží shluku C10. Způsobují to nekvalitní data s nezřetelným trendem.

Vybrané regresní funkce z kroku S2 postupují k následnému testování v S3 s využitím reprezentantů R2. Testování ukazuje, zda jsou vybrané funkce vhodné pro všechny prvky shluku. Krok S2 zahrnuje 10 reprezentantů z celkových 2472 prvků, S3 již obsahuje 43 z 2472 prvků. Tento přístup výrazně sníží časovou náročnost procesu.

Testování (S3) a výsledky

Následuje zhodnocení modelu na základě zvoleného kritéria, tj. podle dat v roce 2015. Jak uvádí text výše, pro každého reprezentanta shluku R1 je vybráno 12 modelů v S2, popř. méně, pokud pro konkrétní data není u některých skupin rovnic zachován monotonní trend. Tato volba je v části S3 zhodnocena indexem determinace pro reprezentanty R1 a R2. Minimální poměr koeficientů determinace R1 a R2 byl nastaven na 0,85, což tedy umožňuje 15% zhoršení pro ostatní prvky shluku. Pokud je toto zhoršení vyšší, z vybraných regresních funkcí není žádná vhodná pro reprezentanta R2 a výpočet se vrací ke kroku S1c) tentokrát s reprezentanty R2 a proces začíná znovu.

Přínos zvoleného přístupu založeného na shlukové analýze zajišťuje vhodně zvolené regresní funkce pro většinu prvků z každého shluku. Testování probíhá na datových sadách reprezentantů jednotlivých podshluků R2, jak ukazuje obr. 6.2 1b). Na této myšlence je založeno také vyhodnocení volby modelů. Vstupní data pro testování zahrnují pouze roky 2009–2014, rok 2015 poslouží k porovnání modelu a skutečných dat. Přičemž uvažovány jsou vždy takové tvary modelů, které přísluší ke shluku (S2), kterému testovací data náleží. Každý další prvek shluku má tedy možnost volby z vybraných modelů. Odhad produkce odpadu oproti datům v roce 2015 dosahuje přibližně v 75 % reprezentantů R2 chyby do 15 % a přibližně polovina reprezentantů má chybu do 5 %. Nejkvalitnější odhady jsou dosaženy v rámci shluků C2 a C4. V těchto případech nelze shledat pravidelnost pro chybu modelu, chyba je pravděpodobně dána oscilací dat okolo trendu. Pro přibližně 9 % reprezentantů R2 překročí odhad produkce chybu 30 %. To může být způsobeno změnou trendu, což není schopna analýza trendu reflektovat, nebo přítomností odlehlých bodů. Této problematice budou věnovány následující kapitoly.

Algorithm 1

1 **S1a)**
2 **Calculate** distance matrix from normalised differencies
3 **S1b)**
4 **Do** hierarchical cluster analysis using Ward's method
5 Choose number of clusters based on criterion $\Rightarrow p$ representatives of type
6 **S1c)**
7 **for all** clusters **do**
8 **Calculate** Silhouettes
9 Choose number of subclusters k based on Table 6.3
10 **Do** cluster analysis using k-means methon $\Rightarrow k$ representatives of type $R2$
11 **end for**
12 **S2**

for all representatives of type $R1$ **do**
    Do function selection using TableCurve 2D
end for
S3
for all representatives p of type $R1$ **do**
    for all respective representatives k of type $R2$ **do**
        Calculate coefficients of determination **I** for $R1_p$ and $R2_k$
        if $\frac{I_{R2_k}}{I_{R1_p}} \leq 0.85$ **and** more then one time series in cluster **then**
            GoTo S1c) with cluster of $R2_k$
        end if
    end for
end for

Výsledkem navrženého algoritmu je seznam tvarů regresních funkcí (6.3 - 6.12), které se jeví jako nejvhodnější pro zpracovávané typy odpadu. Výčet regresních funkcí je zde pro lepší přehlednost uveden v obecném tvaru a nezachovává tak značení z kapitoly 5. Hodnota y udává závisle proměnnou, x je nezávislá proměnná, regresní koeficienty jsou a, b, c, d, e.

$$y = a + bx^c, \tag{6.3}$$

$$y = a + b\exp(-x), \tag{6.4}$$

$$\ln(y) = a + b\exp(-x), \tag{6.5}$$

$$y^{-1} = a + b\exp(-x), \tag{6.6}$$

$$y^{0,5} = a + b\exp(-x), \tag{6.7}$$

$$y = a + \frac{b}{1 + \exp\left(\frac{-(x-c)}{d}\right)}, \tag{6.8}$$

$$y = a + \frac{b}{1 + e(\frac{x-c}{d})^2 \exp\left((1-e)0,5(\frac{x-c}{d})^2\right)}, \tag{6.9}$$

$$y = a + b\frac{(\arctan(\frac{x-c}{d}) + \frac{\pi}{2})}{\pi}, \tag{6.10}$$

$$y = a + b\exp\left(-\exp\left(\frac{-(x - d\ln(\ln(2)) - c)}{d}\right)\right), \tag{6.11}$$

$$y = a + \frac{b}{\left(1 + \frac{(x-c)^2}{d^2 e}\right)^{\frac{d}{2}+\frac{1}{2}}}. \tag{6.12}$$

Jak ukazuje tab. 6.1, shluková analýza nevytvořila shluky, které by byly významněji zastoupeny pouze některými typy odpadů, jako to lze pozorovat v [56]. Výběr regresních funkcí z tohoto důvodu není specifikován pro různé odpady, ale pro veškeré

skupiny odpadů a území jsou připuštěny všechny varianty regresních funkcí 6.3 - 6.12.

Obecně nelze definovat univerzální pravidlo pro popis trendů v produkci odpadu. Extrapolace by měla být vždy provedena ve všech uzlech (ORP, kraje, ČR) pro všechny uzvažované typy odpadu. K tomu je třeba odhadnout tvar regresní funkce, což může být časově náročné. Časové požadavky analýzy mohou být značně sníženy způsobem, který byl popsán v této kapitole.

6.1.2 Nastavení počátečních odhadů

Z vybraných tvarů regresních funkcí 6.3 - 6.12 lze snadno pozorovat, že se jedná o případ nelineární regrese 2.2. Bohužel minimum součtu kvadrátů chyb nelze určit explicitně, jako v případě lineární regrese. Využívají se tak iterativní numerické postupy, které vyžadují nastavení počátečních hodnot parametrů. Kvalitní počáteční odhady napomáhají nalézt řešení rychleji a spolehlivěji. [59] a [60] navrhují následující možnosti pro volbu těchto počátečních odhadů:

- Jak již bylo uvedeno dříve, často se jako nelineární regresní model používá nějaká fyzikální nebo empirická závislost. Pokud mají parametry přesný význam, lze pro počáteční hodnoty využít znalosti z podobných experimentů.
- U linearizovatelných modelů lze odhad parametrů získat z lineární regrese. Tato metoda je vhodná pro získání počátečních odhadů a zjištění odchylek od linearity.
- V případě, že neexistuje žádný vhodný způsob pro volbu startovacích hodnot, je doporučováno použití vyhledávání v mřížce. Toto prohledávání mřížky lze provést generováním rozsáhlého pokrytí možných hodnot parametrů a jejich kombinací a vyhodnocením modelu v každé z těchto možností. Jako výchozí odhady parametrů se potom používá taková kombinace, která vede k nejlepšímu modelu ve smyslu nejmenšího součtu kvadrátů chyb [61].

Parametr a ve všech regresních rovnicích 6.3 - 6.12 představuje absolutní člen, který umožňuje posun modelu po ose y. V modelu je tomuto parametru dán význam, který lze využít pro nastavení počátečních hodnot. Pro parametry a je počáteční hodnota dána produkcí odpadu v posledním sledovaném období, pro data v této práci se tedy jedná o rok 2015.

Vybrané tvary regresních funkcí mají zcela obecný tvar. Pro nastavení počátečních hodnot ostatních regresních koeficientů v této práci tak nelze užít význam těchto parametrů ani linearizaci modelů. Variantou je tedy vytvoření mřížky možných hodnot parametrů. I zde je nutné nastavit krajní hodnoty pro všechny regresní parametry, které budou součástí této mřížky a mohou se tak stát startovacími body pro nelineární regresi.

Cennou informaci o přibližných hodnotách regresních koeficientů lze získat z výpočtů modelů s využitím softwaru TableCurve 2D. Když byl v předchozí části 6.1.1 tento software využit pro volbu regresních funkcí, byly současně zaznamenány hodnoty odhadů regresních parametrů pro všechny vybrané regresní funkce a testované časové řady (reprezentanty R1 a R2). Z těchto informací lze pozorovat v jakých hodnotách se pohybují regresní koeficienty pro jednotlivé regresní funkce a reprezentanty různých průběhů časových řad.

Přes toto významné snížení možných nastavení regresních koeficientů je mřížka startovacích hodnot velmi rozsáhlá i s ohledem na to, že některé regresní funkce mají až pět parametrů. Pro pohyb mezi přípustnými startovacími hodnotami je tedy využito opakovaného generování náhodných kombinací regresních parametrů z rovnoměrného rozdělení:

$$b_k \sim Ro(\min_{b,k}, \max_{b,k}), \quad \forall k \in K, \tag{6.13}$$

kde b_k je počáteční odhad parametru b regresní rovnice s indexem k. Interval rovnoměrného rozdělení určuje minimální a maximální hodnota příslušného koeficientu pro reprezentanty R1 a R2 získané výpočtem v softwaru Table Curve 2D. Ostatní parametry c, d, e se generují stejným způsobem. Při volbě počátečních odhadů je navíc třeba dbát na omezení softwaru GAMS (General Algebraic Modeling System), ve kterém probíhá realizace výpočtu. Interval pro generování některých startovacích hodnot je nutné zúžit s cílem zamezení generování příliš vysokých nebo nízkých hodnot, což vede k přetečení proměnných.

Startovací hodnoty jsou z rovnoměrného rozdělení generovány opakovaně. Zachovává se taková kombinace výchozích hodnot, která vede k nejnižší hodnotě účelové funkce, tedy součtu kvadrátů chyb pro příslušnou časovou řadu dat.

6.1.3 Zachování monotonního trendu a volba modelu

Zvolené funkce 6.3 - 6.12 nemají obecně monotonní průběh. Návrh modelu v podmínce 5.7 požaduje monotónnost funkce popisující trend v datech a za tímto účelem je zavedena binární proměnná $\alpha_{s,t}$. S touto podmínkou přechází úloha nelineárního programování na smíšenou celočíselnou úlohu nelineárního programování, což s sebou nese komplikace z hlediska řešitelnosti.

V tomto případě je tedy přistoupeno ke zcela triviálnímu opatření. Regresní analýza historických dat byla provedena pro všechny vybrané tvary regresních funkcí 6.3 - 6.12. Přičemž jako model trendu v datech příslušné časové řady je uvažován takový regresní model, který dosahuje nejkvalitnějšího proložení historických dat, vyjádřeného prostřednictvím koeficientu determinace. A to za podmínky, že je současně zachován monotónní průběh trendu po sledované období dané množinou I.

6.2 Identifikace odlehlých hodnot

Povinnost evidence produkce a zpracování odpadu upravuje zákon [62], tato povinnost se vztahuje na původce a oprávněné osoby, které s odpady nakládají. Data shromažďuje ISOH, v současné době je správou této databáze pověřena CENIA. Evidence je vedena za každou provozovnu a za každý druh odpadu. Informace se tedy shromažďují od ohlašovatelů, z obecních úřadů ORP, krajských úřadů, MŽP a z Centrální ohlašovny MŽP. Data probíhají kontrolou ORP a poté jsou exportována na MŽP a CENIA, kde probíhá další (celorepubliková) kontrola [27].

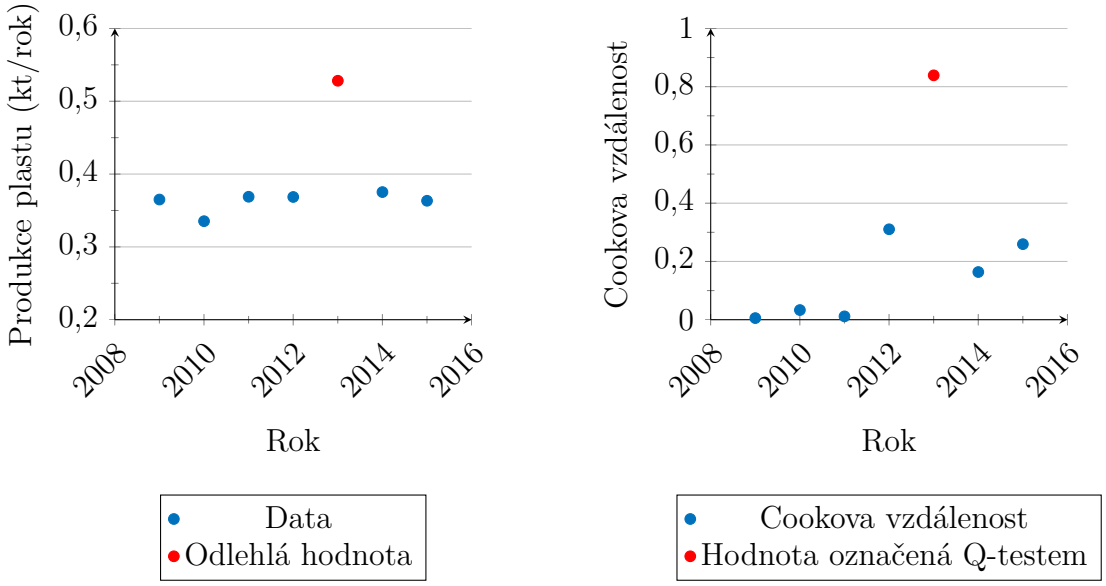
Ačkoliv evidovaná data probíhají vícekrokovou kontrolou, nelze eliminovat veškeré chyby vzniklé při sběru dat. Mezi chyby, kterých se respondenti nejčastěji dopouštějí, patří nesprávné zadání kódu (území, nakládání, typu odpadu), evidence množství odpadu ve špatných jednotkách. Případně se jedná o výkyvy v produkci, o kterých není možné zjistit více informací např. zpětná těžba starých ekologických zátěží.

Přítomnost odlehlých a extrémních hodnot může způsobit zkreslení výsledků regresní analýzy. Data zpracovávaná v této práci (kap. 1.1) jsou dána vždy v určitých letech pevně daného rozsahu (2009 - 2015), není tedy možné, aby se v datech vyskytovaly extrémní hodnoty. Tato kapitola bude věnována identifikaci pouze odlehlých hodnot, tedy výkyvům v množství vyprodukovaného odpadu.

K ověření přítomnosti vlivných bodů poslouží Cookova vzdálenost 2.17. Tuto míru lze využít jak u lineární tak i nelineární regrese, viz [35]. Jak zmiňuje [38], vhodnou interpretací Cookovy vzdálenosti je individuální posouzení každé datové sady, namísto nastavení pevně dané kritické hodnoty, která bývá pro nelineární regresi nejčastěji rovna 1. Dean-Dixonův test v podobě Q_n je aplikován na Cookovy vzdálenosti každé časové řady, čímž slouží jako kritérium pro detekci odlehlosti Cookových vzdáleností. Pokud Q_n převyšuje kritickou hodnotu $Q_{0,05}$ je bod s nejvyšší Cookovou vzdáleností označen jako odlehlý na hladině významnosti $\alpha = 0,05$ a pro další výpočty odstraněn.

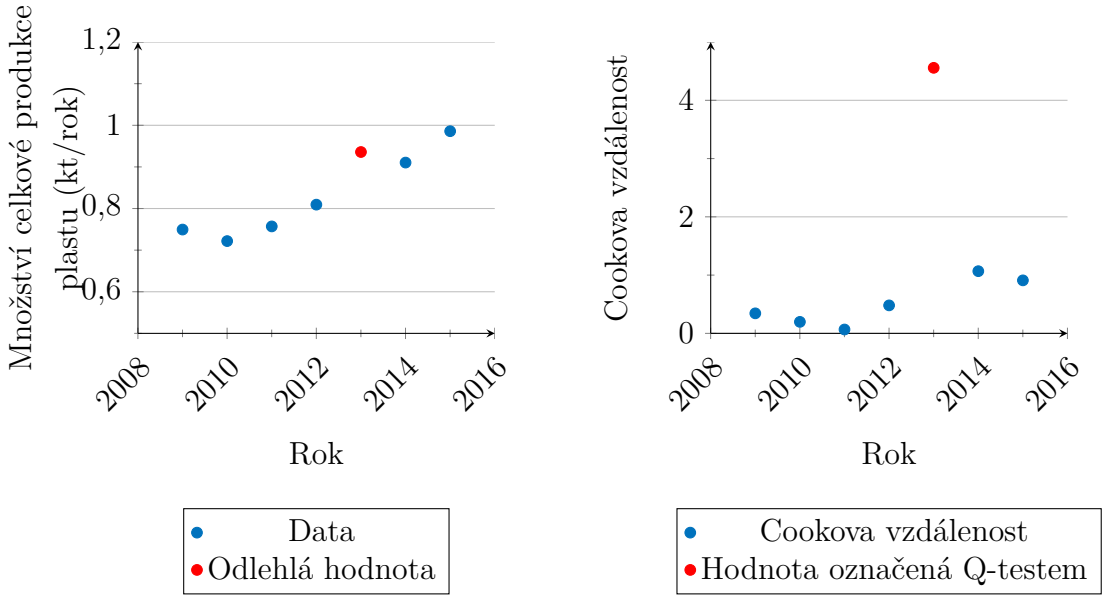
Nutné je podotknout, že vlivné body nemusí být nutně vadné hodnoty. Případy, které mohou nastat, jsou zobrazeny v následujících grafech.

Obr. 6.5 ukazuje situaci, kdy je s pomocí Cookovy vzdálenosti úspěšně detekována odlehlá hodnota v roce 2013 a následně tedy odstraněna, aby se zamezilo jejímu negativnímu vlivu na výsledný model. Lze vidět, že pokud by byla kritická hodnota Cookovy vzdálenosti nastavena pro všechny případy rovna 1, jak tomu v řadě aplikací bývá [35], tato odlehlá hodnota by nebyla odhalena. Q-test pro nejvyšší hodnotu Cookovy vzdálenosti této časové řady dosahuje $Q_7 = 0,648 > Q_{0,05} = 0,507$.



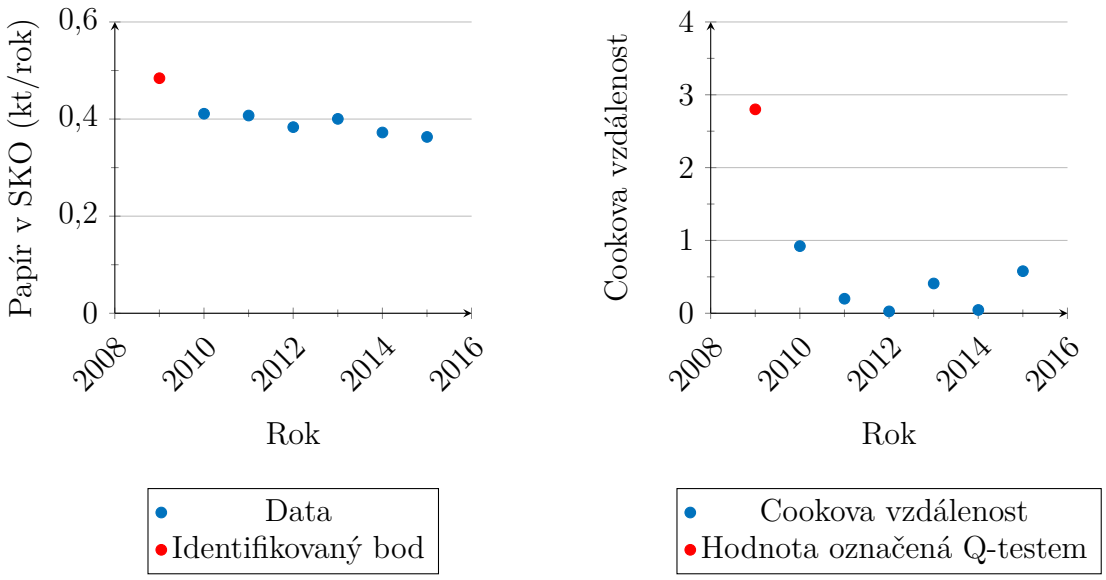
Obr. 6.5: Úspěšná detekce odlehlého bodu na základě Cookovy vzdálenosti

Rozdíl mezi modelem na základě kompletních dat a po vynechání některého bodu je normován počtem parametrů příslušné rovnice a rozptylem reziduí kompletního modelu. To vede k případu, který je zobrazen na obr. 6.6. U velmi kvalitních dat sebemenší odchylka od trendu navýší Cookovu vzdálenost tak, že dojde k detekování odlehlého bodu. S odstraněním takovýchto hodnot lze souhlasit, i zde dochází k ovlivnění jinak kvalitních dat.



Obr. 6.6: Vyloučení bodu na základě Cookovy vzdálenosti u kvalitních dat

Pozornost je třeba věnovat okrajovým bodům datové sady, které mají výrazný vliv na podobu regresního modelu, jak ukazuje Cookova vzdálenost ilustrována na obr. 6.7. Produkce v počátečním roce (2009) je zřejmě součástí klesajícího trendu papíru v SKO a nejedná se o chybu v datech. Na model má však natolik vysoký vliv, že její Cookova vzdálenost převyšuje vzdálenosti v ostatních letech a je tak detekována Q-testem, což poukazuje na odlehlou hodnotu v datech. Celkem 25,41 % počátečních bodů je Cookovou vzdáleností identifikováno jako odlehlé a totéž platí pro 15,76 % koncových bodů časové řady. Cílem je dostat co možná nejkvalitnější data, zároveň je třeba dbát opatrnosti při vypouštění hodnot, protože by mohlo dojít ke ztrátě informace o klesající podobě trendu.



Obr. 6.7: Vliv krajních bodů na regresní model

Metodika byla doplněna o analýzu reziduí v krajních bodech časové řady. Pro aditivní model jsou rezidua definována vztahem 2.5. Podle [31] lze k analýze reziduí užívat názorného grafického zobrazení. Hodnota je podezřelá z odlehlosti, pokud se reziduum v tomto bodě výrazně odklání od reziduí v bodech ostatních. Za předpokladu normálního rozdělení reziduí $\widehat{\varepsilon}_i$, $\widehat{\varepsilon}_i \sim N(0, \sigma^2)$, leží 95,4 % těchto reziduí v intervalu $(-2\sigma, 2\sigma)$, kde σ je směrodatná odchylka reziduí. Pokud reziduum krajního bodu časové řady leží mimo zmíněný interval a navíc Cookova vzdálenost přesahuje kritickou hodnotu, je tento krajní bod označen jako odlehlý.

Jak ukazují grafy této kapitoly, identifikace odlehlých hodnot s využitím Cookovy vzdálenosti může být užitečným nástrojem. Je však třeba s ním zacházet s opatrností, kde zejména v krajních bodech časové řady by mohlo docházet k odstraňování kvalitních dat. Cílem této analýzy je odstranění vadných hodnot, ale nikoliv na úkor ztrát informací, které by mohly vzejít z nesprávné identifikace odlehlých hodnot.

Odlehlé hodnoty uvnitř časové řady jsou identifikovány s využitím pouze Cookovy vzdálenosti v podobě 2.17, v krajních bodech je toto kritérium rozšířeno. Po identifikaci vlivného bodu byla porovnána rezidua a hodnoty dat zachovány, pokud se rezidua výrazně nelišila od reziduí v ostatních bodech.

6.3 Shrnutí

Řešení matematického modelu ve formě dekompozice jistě vede k chybě ve výpočtech, které se při tomto přístupu dopouštíme. Pro část věnované agregaci, tento odklon není významný, protože množina všech možných agregací je omezena na případy, které produkují co nejkvalitnější data. Podobná situace nastává i u analýzy trendu. Množina regresních funkcí je zúžena na funkce dosahující nejlepšího modelu dat a tak nelze předpokládat, že by bylo dosaženo významně rozdílných výsledků od globálního modelu. Chyby mohou být ale způsobeny využitím shlukové analýzy pro tento výběr regresních funkcí, kde jsou regresní funkce voleny pouze pro reprezentanty shluků. Identifikace odlehlých bodů pro globální problém a dekomponovanou formu má totožný princip. Dalším odklonem je závěrečná bilance nástrojem Justine. V případě kompletního modelu dochází k této bilanci již ve fázi modelování trendu v datech a model je tak touto bilancí přímo ovlivněn. Tzn. že navržené extrapolace zaručí soulad hodnot v rámci hierarchické struktury. V dekomponované části probíhá bilance až na samotném závěru a extrapolace představují pouze počáteční hodnotu, která je zpřesněna. Kvantifikace chyb, ke kterým dochází při aplikace dekomponované úlohy, je námětem pro další vývoj a zkvalitnění prognózování.

7 PŘÍPADOVÁ STUDIE

7.1 Formulace úlohy a její implementace

Následující kapitola se věnuje aplikaci popsané metodiky pro odhad budoucí produkce odpadů na datech KO v ČR 1.1. Přístup realizuje navržený matematický model ve formě dekompozice, tak jak byl uveden v kapitole 6 s postupným řešením dílčích úloh, jak graficky znázorňuje obr. 6.1. Výpočet je implementován v systému GAMS, kde vznikl komplexní nástroj pro prognózování produkce odpadů.

Prognóza KO, jejíž výstupy jsou shrnuty v této kapitole, vzniká pro tři opakovaně generované scénáře počátečních odhadů, viz kap. 6.1.2. Tab. 7.1 shrnuje úlohu z hlediska počtu proměnných a časové náročnosti[1].

	Počet úloh nelin. prog.	Celkový počet regr. parametrů	Časová náročnost (h)
Výběr funkce	79 560	262 548	19,81
Odlehlé body	55 692	32 538	13,86
Výsledný model	3 054	12 645	0,76

Tab. 7.1: Charakteristika přístupu z hlediska implementace

7.1.1 Agregace dat

Z důvodu interpretovatelnosti výsledků jsou uvažovány agregace území na základě legislativního členění ČR na kraje a ORP. Právě území takovéhoto tradičního dělení bývají nejčastěji cílem prognózy. Hierarchická struktura územních celků ČR je tedy členěna na tři stupně této struktury.

Agregace typů odpadů je založena na principu popsaném v kapitole 1.1 a to opět s ohledem na interpretaci výsledků pro tříděné složky a složení SKO. V zájmu prognózy jsou obvykle právě tříděné složky odpadu, nikoliv jednotlivá katalogová čísla. Důležitou informaci z pohledu separace také poskytuje odhad složení SKO a to z důvodu potenciálu pro nárůst vytříděných KO.

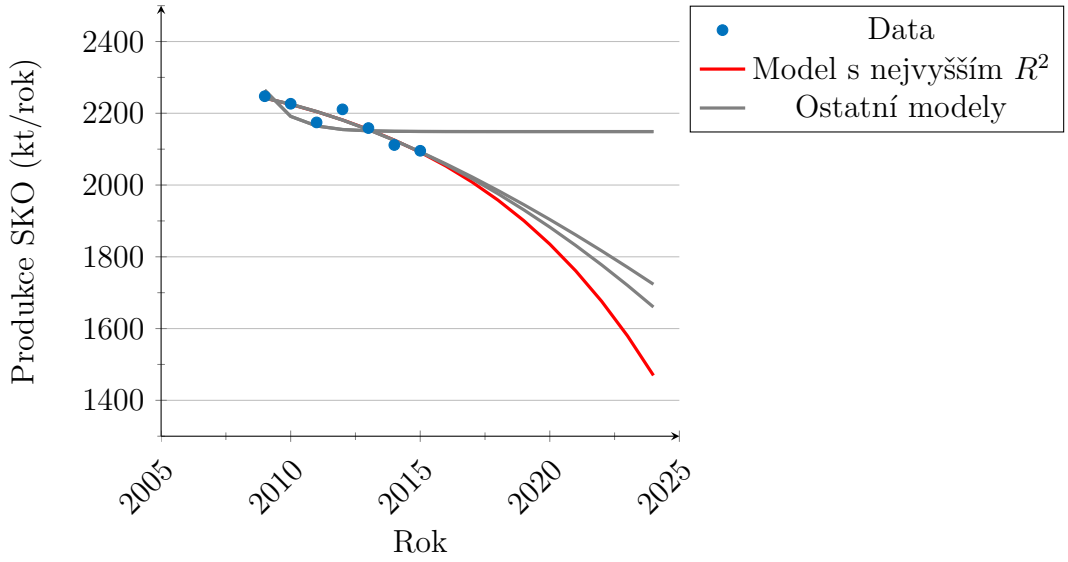
[1]Výpočty probíhaly na počítači s těmito parametry:
Procesor: Intel(R)Core(TM)i7-5500U CPU @2.40GHz 2.40 GHz
Nainstalovaná paměť: 8,00GB
Typ systému: 64bitový operační systém pro platformu x64

7.1.2 Analýza trendu

V této části je s využitím regresní analýzy vytvořen model trendu v historických datech (způsobem popsaným v kapitole 6.1) pro všechny uvažované časové řady. Problematika je řešena celkem pro 221 územních celků na třech úrovních hierarchické struktury (206 ORP, 14 krajů a ČR) a 12 typů odpadu. Jedná se tedy o 2652 časových řad, pro zde zpracovávaná data bylo vybráno celkem 10 typů regresních funkcí 6.3 - 6.12. Celkem je tedy řešeno 26520 samostatných úloh nelineárního programování a tento počet se násobí s vyšším množstvím generovaných startovacích hodnot.

Pro popis trendu v datech je pro každou časovou řadu následně využit jediný tvar regresní funkce z deseti možných, který dosahuje nejvyšší hodnoty koeficientu determinace a zároveň zachovává monotonní průběh ve sledovaném období I. Pro veškeré výpočty v této aplikaci jsou počáteční odhady generovány třikrát a zachovává se varianta vedoucí na nejnižší hodnotu účelové funkce.

Na obr. 7.1 jsou zobrazena data o produkci SKO a trend v těchto datech na základě různých tvarů regresních funkcí, přičemž některé z nich se překrývají. Model vyznačený červenou barvou dosáhl nejvyšší hodnoty koeficientu determinace R^2 ($R^2 = 0,898$) a následně se v případě časové řady SKO v ČR pracovalo pouze s tímto trendem.



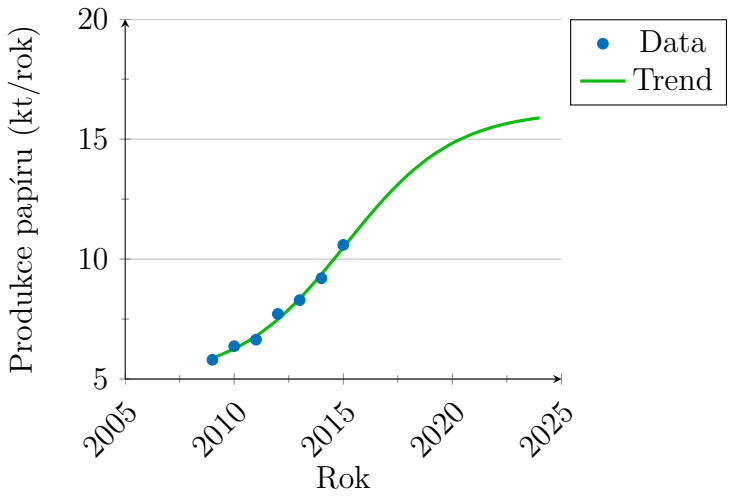
Obr. 7.1: Volba trendu

Procentuální využití jednotlivých regresních funkcí pro trend v historických datech ze všech časových řad shrnuje tab. 7.2. Velmi výrazně převyšuje ostatní funkce tvar (6.8). Jako příklad je uvedent trend v datech papíru modelovaný právě touto funkcí, viz obr. 7.2, koeficient determinace je pro tento případ 0,83. Jedná se funkci,

která se řadí k tzv. S-křivkám, práce [23] také zmiňovala tento typ křivek jako nejvhodnější model trendu v datech pro SKO a jeho složky.

Funkce	6.3	6.4	6.5	6.6	6.7	6.8	6.9	6.10	6.11	6.12
Množství časových řad [%]	7,39	0,15	0,28	0,11	0,11	75,20	2,07	0,34	0,11	14,44

Tab. 7.2: Procentuální využití jednotlivých tvarů regresních funkcí



Obr. 7.2: Trend modelovaný S-křivkou pro konkrétní komoditu PAP

Omezení minimální a maximální produkce pro model trendu v datech, jako v podmínce 5.6, je pro pouhé zamezení extrémnímu poklesu nebo růstu nastaveno na 50 % z nejnižší produkce v období J pro dolní mez. Maximální hodnota produkce dovoluje 50% nárůst oproti nejvyšší produkci z období J.[1]

7.1.3 Identifikace odlehlých pozorování

Metodika pro detekci odlehlých hodnot byla popsána v části 6.2 tohoto textu. Výpočet Cookovy vzdálenosti 2.17 je založen na porovnání regresního modelu pro kompletní data a modelu pro případ s vynechaným bodem. Na tomto místě je tedy třeba určit aproximaci historických dat s postupným vynecháváním jednotlivých bodů. Pro výpočet Cookovy vzdálenosti je už využita pouze jediná regresní funkce vybraná v části 7.1.2.

Následuje tedy další fáze tvorby regresních modelů trendu pro 2652 časových řad, kde je vždy vynechán jeden ze sedmi bodů datové sady. V této části je tedy

[1] Vychází z praktických zkušeností odborníků z ÚPI.

řešeno 18564 úloh nelineárního programování. Zde je již znám tvar regresní funkce, je využita tedy pouze jedna rovnice a počet hledaných regresních parametrů se tak pohybuje v rozmezí od 2 do 5 pro každou úlohu.

I zde je řešena otázka monotonie trendu a nastavení počátečních odhadů parametrů. Pro tuto část regresní analýzy s vynecháváním hodnot je výhodou informace o výsledcích výpočtu pro kompletní data. Pro zajištění monotonního průběhu trendu je přidána jedna z podmínek 7.1, 7.2. V závislosti na tom, zda byl trend pro kompletní data rostoucí nebo klesající je rozdíl vždy dvou po sobě jdoucích odhadů větší nebo menší než nula. Je rovněž připuštěna rovnost nule, tzn. možností je i varianta konstantního průběhu trendu nebo jeho části.

$$m_{i+1,s,t} - m_{i,s,t} \leq 0, \forall s \in S, \forall t \in T, \forall i = 1, .., |I| - 1, \tag{7.1}$$

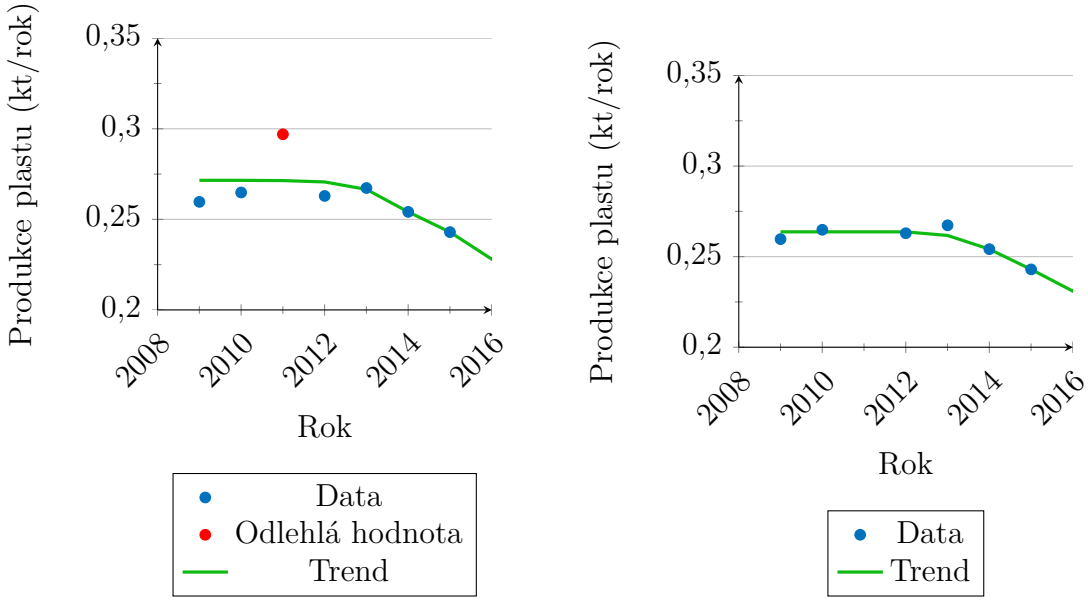
$$m_{i+1,s,t} - m_{i,s,t} \geq 0, \forall s \in S, \forall t \in T, \forall i = 1, .., |I| - 1. \tag{7.2}$$

Za předpokladu kvalitních dat se regresní model s vynechaným bodem výrazně neodkloní od regresního modelu pro kompletní datovou sadu, vhodnými počátečními odhady parametrů pro vynechávané body se stávají odhady parametrů regresní funkce pro kompletní data. Není tedy nutné v této části regresní analýzy opakovaně generovat počáteční hodnoty.

Na základě Cookovy vzdálenosti, která je v krajních hodnotách časové řady doplněna o analýzu reziduí, metodika navrhuje odstranění 4,80 % bodů ze všech dostupných dat. U 66,93 % časových řad není třeba odstraňovat žádný z bodů příslušné datové sady, u zbývajících časových řad tedy dochází k identifikaci alespoň jednoho odlehlého bodu.

7.1.4 Model trendu bez odlehlých hodnot

V případě, že na základě přístupu popsaného v kap. 6.2, je detekován odlehlý bod, pro výpočet trendu v datech se odstraňuje bez náhrady. Přibližně u 67 % časových řad není detekována žádná odlehlá hodnota. Obr. 7.3 ilustruje změnu trendu po vynechání odlehlého bodu v roce 2013 pro data o produkci plastu.



(a) Trend za přítomnosti odlehlé hodnoty

(b) Trend po odstranění odlehlé hodnoty

Obr. 7.3: Ukázka trendu v datech pro PL v konkrétním ORP

7.1.5 Závěrečná bilance nástrojem Justine

Posledním krokem popsané metodiky na základě schématu obr. 6.1 je bilance v nástroji Justine [54] ve sledovaném roce. V této aplikace je pozornost věnována roku 2024, jedná se o období, kdy má vzejít v platnost zákaz skládkování SKO. Z tohoto důvodu je v současné době prognóza produkce odpadu v roce 2024 zásadní.

Úprava Justine zajišťuje soulad mezi predikovanými hodnotami v požadovaném roce s ohledem na hierarchické uspořádání území a katalogových čísel odpadu s požadavkem na minimální odklon od modelu trendu v datech.

Potenciální nárůst separace materiálově využitelných složek je dán obsahem těchto složek v SKO. Efektivitu třídění lze podle [30] kvantifikovat tzv. mírou separace (MS), která udává procentuální množství vyseparovaného odpadu. Současně tedy podává informaci o možném zvýšení separace odpadu.

$$MS = \frac{m_{vytříděno}}{m_{vytříděno} + m_{SKO}}, \tag{7.3}$$

kde $m_{vytříděno}$ vyjadřuje množství vytříděného odpadu vybraného druhu a m_{SKO} je množství tohoto odpadu, které zbylo v SKO.

Kromě podmínek zachovávající hierarchickou strukturu dat podle 5.23 a 5.24 doplňuje závěrečnou bilanci také omezení míry separace. Nastavení maximální MS je inspirováno zahraničními zkušenostmi. Na základě studie současné produkce a kvality separace odpadu v Německu [63] je stanovena hranice MS do roku 2024 v ČR pro tři typy zástavby. S ohledem na dosavadní vývoj OH v ČR nelze předpokládat

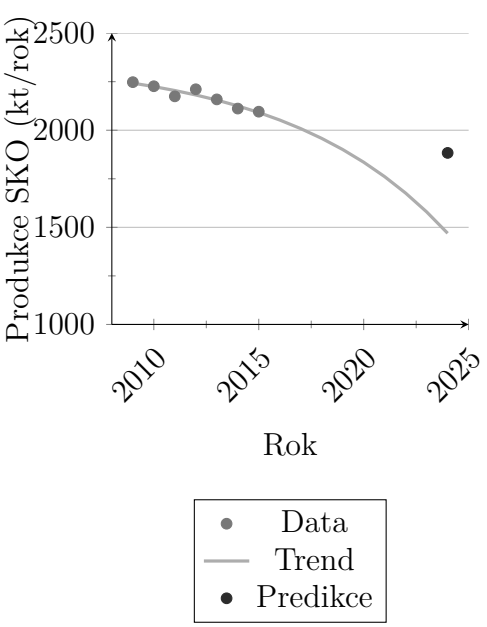
tak dramatický vzestup separovatelnosti odpadu, který by překročil aktuální MS v Německu a proto je možné tato data pokládat za horní hranici MS. Omezení pro různé typy zástavby shrnuje tab. 7.3.

Typ zástavby	Maximální MS [%]		
	papír	plast	sklo
Městká zástavba	70	52	73
Přechodná zástavba	86	64	89
Vesnická zástavba	93	73	90

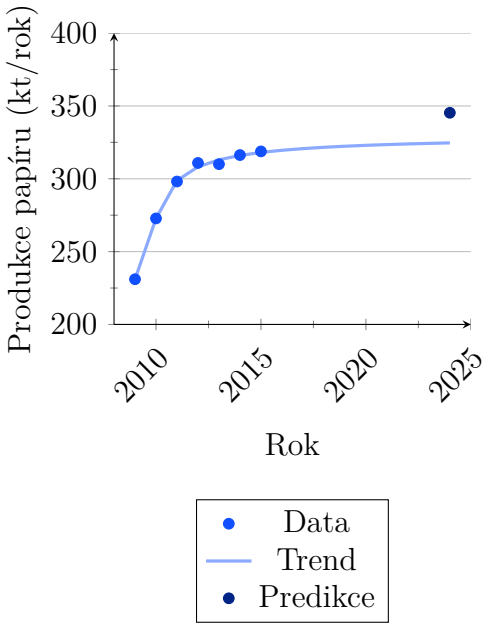
Tab. 7.3: Omezení maximální MS na základě typu zástavby

Výsledné hodnoty predikce v roce 2024 pro separovaný odpad a SKO včetně korigovaných hodnot s využitím Justine lze vidět na obr. 7.4 v rámci celkových výsledků označené jako bod "Predikce".

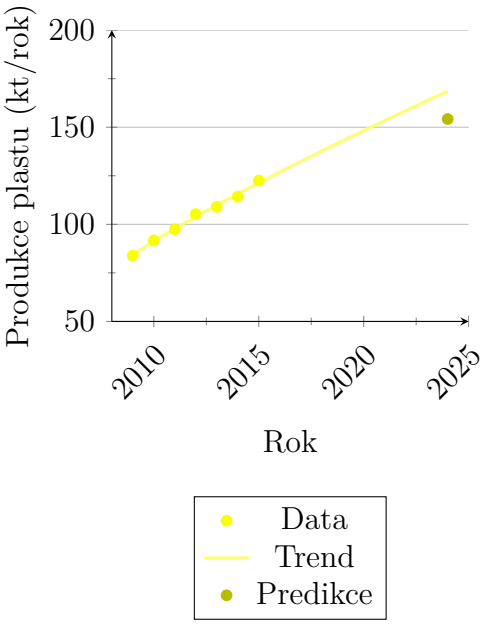
7.2 Výsledky



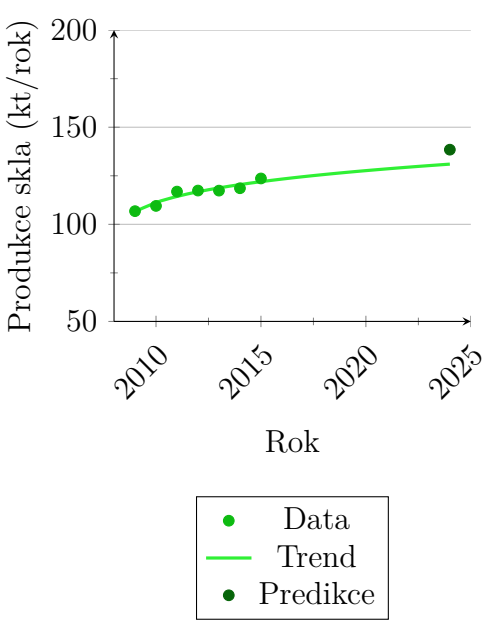
(a) Prognóza produkce SKO



(b) Prognóza produkce papíru



(c) Prognóza produkce plastu

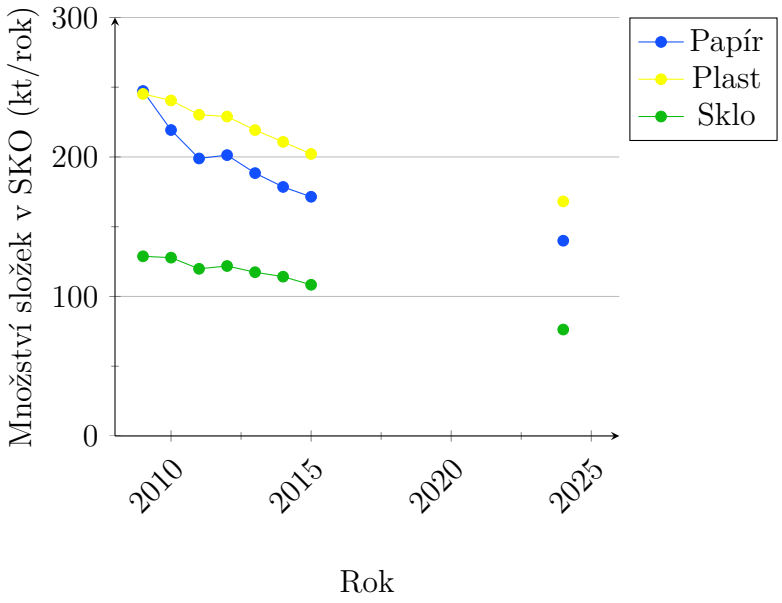


(d) Prognóza produkce skla

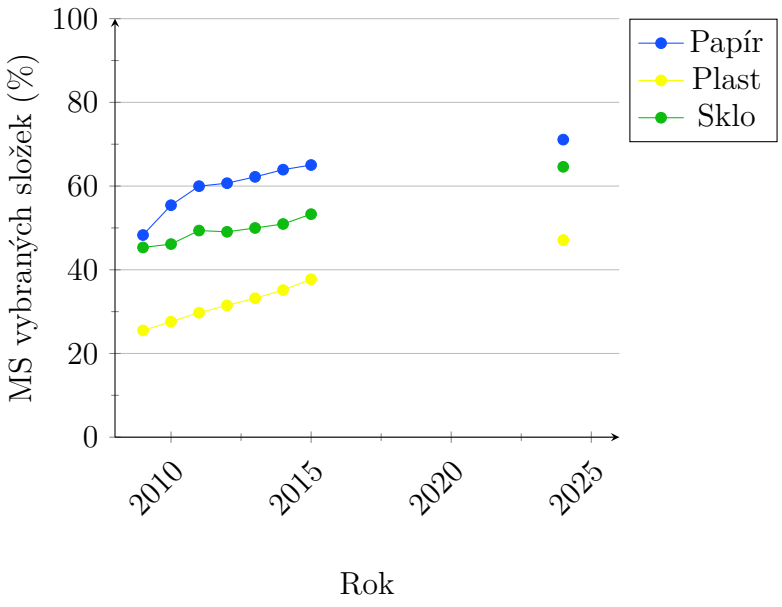
Obr. 7.4: Ukázka výsledků prognózy pro SKO a tříděné složky v ČR

Prognózu produkce SKO a separovaných složek odpadu (papír, plast, sklo) v ČR zobrazuje obr. 7.4. Pro názornost jsou v grafickém výstupu znázorněna data o produkci a extrapolace těchto dat. Rovněž lze sledovat korigované hodnoty pro predikci v roce 2024, které se především u SKO a produkce plastu odklání od trendu k vyšší produkci. Analýza trendu nezaručuje zachování vztahů agregovaných dat v hierarchické struktuře. Této vlastnosti je dosaženo závěrečnou bilancí, kdy je hodnota extrapolace ve sledovaném roce zpřesněna [24].

Vývoj množství sledovaných složek v SKO v období 2009 - 2015 ilustruje obr. 7.5. Součástí grafu jsou i body pro odhad množství těchto složek v roce 2024. Do tohoto období lze sledovat pokles všech složek, který je způsoben jak dlouhodobě klesajícím trendem produkce SKO 7.4, tak vyšší MS.



Obr. 7.5: Vývoj sledovaných složek v SKO v letech 2009 - 2015 a predikce v roce 2024



Obr. 7.6: MS v letech 2009 - 2015 a nárůst do roku 2024

Výhled pro nárůst MS do roku 2024 ilustruje obr. 7.6. Současně tyto výstupy nesou informaci o efektivitě třídění a tedy možném zvýšení separace odpadu pro následující období.

Grafy vykreslené v této kapitole jsou ukázkou informací, které vzešly z metodiky popsané v této práci. Na podobném principu lze tvořit závěry pro další skupiny odpadů a územní celky (kraje, ORP).

8 MOŽNOSTI DALŠÍHO VÝVOJE

Aplikace zde popsané metodiky na reálných datech poodkrývá možnosti pro navazující doplnění vzniklého nástroje. Další vývoj této práce by tak mohl vést ke zkvalitnění odhadů budoucí produkce a rozšíření možností využití výstupů.

Doplnění o socio-ekonomická data

V současné podobě modelu je jedinou nezávisle proměnnou průběh produkce odpadu v čase a jedná se tak o velmi zjednodušený případ. Klíčovou úlohu však hrají také dlouhodobé sociální, demografické a ekonomické trendy [64], jejichž dopad na produkci a složení odpadu by měl být zohledněn v dalších fázích vývoje modelu.

Nastavení maximální a minimální produkce

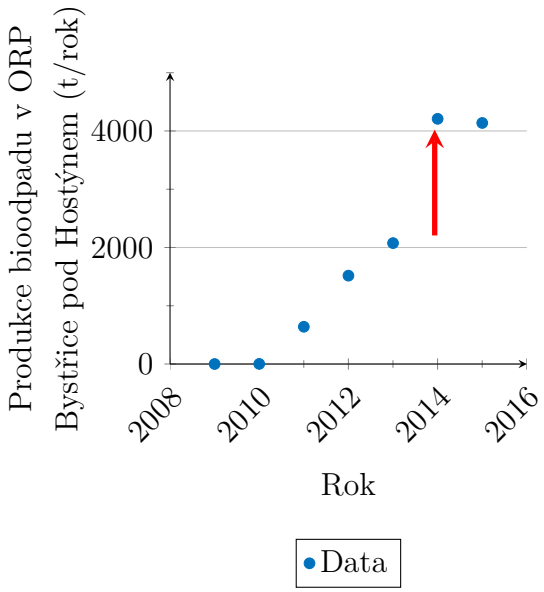
Cílem je stanovení okrajových podmínek pro produkci vybraných odpadů, čímž by se zamezilo poklesu nebo nárůstu trendu mimo reálné meze produkce daného typu odpadu, což je v modelu ošetřeno prostřednictvím podmínek 5.6 a 5.22. Pro potřeby této práce je minimální produkce uvažována jako 50% z nejnižší produkce v období J. Maximální hodnota produkce dovoluje 50% nárůst oproti nejvyšší produkci z období J. Toto opatření je poněkud těžkopádné a vychází pouze z potřeby nastavit trendu v datech hranice. Nabízí však možnosti sofistikovanějšího přístupu, jako může být inspirace zkušenostmi ze zahraničí s výtěžností různých typů odpadu.

Rozšíření aplikace na nižší územní celky

Nejnižším územním celkem pro aplikaci na datech jsou aktuálně ORP. V případě dostupnosti dat na úrovni obcí by bylo možné model doplnit o další část hierarchické struktury a provést testování na nejnižším stupni, kde lze očekávat ještě výraznější variabilitu dat, než je tomu pro ORP. ČR je rozdělena na 206 ORP a asi 6250 obcí. Vzhledem k výpočtovým časům uvedeným v tab.7.1, toto rozšíření by mohlo představovat komplikovaný problém z hlediska implementace a proto je třeba uvažovat uvedené zjednodušení z kap. 6.

Identifikace skokových změn

Obr. 8.1 ilustruje situaci, kdy v roce 2014 došlo k nějaké blíže nespecifikované změně v produkci. Trend se v tomto roce výrazně změnil, bylo by tedy vhodné takovéto zásahy nějakým způsobem zohlednit pro kvalitnější prognózu trendu. Cílem je tedy vytvořit metodiku pro odhalení skokových změn trendu.



Obr. 8.1: Data se skokovou změnou produkce

Rozšíření na další typy odpadu

Současná aplikace se soustředí na vybrané skupiny KO. Identifikace skokových změn z předchozího bodu by přímo souvisela s kvalitnější aplikací např. na datech bioodpadu. V roce 2014 byla přijata novela zákona o odpadech, která obcím stanovila povinnost zajistit pro domácnosti místa pro odkládání bioodpadu. V tomto období lze tedy očekávat skokové změny trendu v produkci.

Scénářový přístup a nastavení váhování

Ze seznamu vhodných regresních funkcí je na základě koeficientu determinace vybrán jediný tvar pro modelování trendu. Takovýto trend popisuje jeden z možných scénářů vývoje produkce odpadu, ke kterému může v průběhu času dojít. Zachování více modelů by popsalo více možných scénářů. Pro jejich implementaci do Justine by bylo třeba vyvinout systém váhování pro bilanci tohoto scénářového řešení.

9 ZÁVĚR

Tato diplomová práce je dílčí součástí problematiky zabývající se odhadem budoucí produkce odpadů. Konkrétněji se věnuje tvorbě predikčních modelů a výsledky jsou vstupem pro nástroj Justine [54]. Model produkce opadu je tvořen na základě historických dat, aktuálně jsou k dispozici data z let 2009 - 2015. Práce se tedy zabývá prognózou na základě krátké časové řady, kdy selhávají tradiční strategie tvorby prognostických modelů.

Po úvodu seznamujícího čtenáře se zpracovávanou problematikou a zavedení potřebného matematického aparátu následuje představení návrhu matematického modelu, kap. 5. Tento model je vystavěn na principu metody nejmenších čtverců pro modelování trendu v historických datech. Model trendu je extrapolován pro následující období, přičemž jedinou nezávisle proměnnou je průběh v čase. V první části modelu (kap. 5.1) jsou data agregována s cílem snížení variability dat, která je typická zejména pro nižší územní jednotky a je tak vybudována hierarchická struktura. Protože každá časová řada dostupných dat, pro kterou je třeba modelovat trend v datech, může mít zcela rozdílný průběh v čase, je v modelu ponechána možnost volby tvaru regresní funkce pro každou časovou řadu. Součástí metodiky je také identifikace odlehlých pozorování, jejichž přítomnost má významný vliv na podobu modelu. Navazuje druhá část modelu (kap. 5.2), která využívá informaci o vybraných regresních funkcích a vynechává body označené jako odlehlé hodnoty. Trend v datech a jeho extrapolace jsou již tvořeny se současnou bilancí ve sledovaném roce s cílem zachování vazeb hierarchické struktury.

Matematický model je úlohou smíšeného celočíselného nelineárního programování a pro jeho řešení je navržen systém dekompozice, kap. 6. Tento přístup postupně řeší jednotlivé kroky v návaznosti za sebou. Otázka agregace dat a závěrečné bilance Justine byla v zájmu autorů již dříve zpracovávaných studií, v této práci je větší prostor věnován analýze trendu a detekce odlehlých bodů.

Analýza trendu (kap. 6.1) zahrnuje návrh metodiky pro volbu regresní funkce pro každou časovou řadu. S cílem snížení časové náročnosti je využita shluková analýza pro seskupení producentů s podobným průběhem v čase a následně se pracuje pouze s reprezentanty těchto skupin. Reprezentantům jsou s využitím softwaru TableCurve 2D vybrány regresní funkce kvalitně modelující trend v datech a tento výběr je následně testován na dalších prvcích každého shluku. Výstupem zmiňované části dekompozice je seznam deseti tvarů regresních funkcí, z nichž lze pro každou časovou řadu z dostupných dat vybrat takovou, která nejlépe modeluje trend v datech z pohledu koeficientu determinace. Protože v případě nelineární optimalizace není zaručeno nalezení globálního minima, jsou počáteční odhady regresních parametrů opakovaně generovány z rovnoměrného rozdělení a zachovává se takové řešení,

které vede na nejnižší hodnotu účelové funkce.

Pro identifikaci vlivných bodů (kap. 6.2) je využita Cookova vzdálenost, což je míra založená na vypouštění bodů z dat a sledování změn modelu, ke kterým po takovém vyřazení bodu dojde. Za vlivný bod je označena taková hodnota, která dosahuje výrazně vyšší Cookovy vzdálenosti, než ostatní body v dané časové řadě. Cookovy vzdálenosti každé časové řady jsou testovány Dean-Dixonovým Q-testem na hladině významnosti $\alpha = 0,05$. Protože krajní body se velmi často jeví jako vlivné, aniž by se jednalo o odlehlé hodnoty, je v krajních bodech zmíněný přístup doplněn o analýzu reziduí.

Matematický model ve formě dekompozice je aplikován v rámci případové studie na data KO v ČR s cílovým rokem 2024, kap. 7. Na základě této aplikace lze tvořit závěry o odhadech produkce v roce 2024 pro jednotlivé typy odpadu, včetně MS na úrovni ORP, krajů a ČR.

LITERATURA

[1] *Zákon č. 185/2001 Sb. Zákon o odpadech a o změně některých dalších zákonů.* In: Sbírka zákonů České republiky. Dostupný také z /www.platnalegislativa.nsf/

[2] MCDONOUGH, William. a Michael BRAUNGART. *Cradle to cradle: remaking the way we make things*. New York: North Point Press, 2002. ISBN 978-0-86547-587-8.

[3] *Novela zákona o odpadech č. 229/2014 Sb.* Zákon, kterým se mění zákon č. 185/2001 Sb., o odpadech a o změně některých dalších zákonů, ve znění pozdějších předpisů.

[4] ŠOMPLÁK, Radovan, Tomáš FERDAN, Martin PAVLAS a Pavel POPELA. Waste-to-energy facility planning under uncertain circumstances. *Applied Thermal Engineering* [online]. 2013, 61(1), 106-114 [cit. 2018-01-21]. DOI: 10.1016/j.applthermaleng.2013.04.003.

[5] ŠOMPLÁK, Radovan a Martin PAVLAS. Potenciál výroby energie z odpadů v případě zákazu skládkování. *Waste forum*. 2014, (2), 95-104. ISSN 1804- 0195.

[6] FERDAN, Tomáš, Radovan ŠOMPLÁK, Lenka ZAVÍRALOVÁ, Martin PAVLAS a Lukáš FRÝBA. A Waste-to-Energy Project: A Complex Approach towards the Assessment of Investment Risks. *Applied Thermal Engineering*. 2015, 89(1), 1127-1136. ISSN 1359- 4311.

[7] BEIGL, Peter, Sandra LEBERSORGER a Stefan SALHOFER. Modelling municipal solid waste generation: A review. *Waste Management*. 2008, 28(1), 200-214. DOI: 10.1016/j.wasman.2006.12.011.

[8] KOLEKAR, K.A., T. HAZRA a S.N. CHAKRABARTY. A Review on Prediction of Municipal Solid Waste Generation Models: A review. *Procedia Environmental Sciences*. 2016, 35(1), 238-244. DOI: 10.1016/j.proenv.2016.07.087.

[9] ORSONI, A., N. V. KARADIMAS a Vassili LOUMOS. Municipal Solid Waste Generation Modelling Based On Fuzzy Logic: A case study of Beijing. *ECMS 2006 Proceedings edited by: W. Borutzky, A. Orsoni, R. Zobel.* ECMS, 2006, 2006-05-28, 409(20), 309-314. DOI: 10.7148/2006-0309. ISBN 0955301807.

[10] LOZANO-OLVERA, Gabriela, Sara OJEDA-BENÍTEZ, Juan Ramón CASTRO-RODRÍGUEZ, Miguel BRAVO-ZANOGUERA a Antonio RODRÍGUEZ-DIAZ. Identification of waste packaging profiles using fuzzy logic: A case study of Beijing. *Resources, Conservation and Recycling.* ECMS,

2008, 2006-05-28, 52(8-9), 1022-1030. DOI: 10.1016/j.resconrec.2008.03.008. ISBN 0955301807.

[11] THANH, Nguyen Phuc, Yasuhiro MATSUI a Takeshi FUJIWARA. Household solid waste generation and characteristic in a Mekong Delta city, Vietnam: note II. *Procedia Environmental Sciences.* 2016, 1968, 35(1), 238-244. DOI: 10.1016/j.jenvman.2010.06.016. ISBN 10.1016/j.jenvman.2010.06.016.

[12] LEBERSORGER, S. a P. BEIGL. Municipal solid waste generation in municipalities: Quantifying impacts of household structure, commercial waste and domestic fuel. *Waste Management* [online]. 2011, 31(9-10), 1907-1915 [cit. 2018-01-24]. DOI: 10.1016/j.wasman.2011.05.016.

[13] LI, Zhen-shan, Hui-zhen FU a Xiao-yan QU. Estimating municipal solid waste generation by different activities and various resident groups: A case study of Beijing. *Science of The Total Environment.* 2011, 1968, 409(20), 4406-4414. DOI: 10.1016/j.scitotenv.2011.07.018. ISBN 10.1016/j.jenvman.2010.06.016.

[14] ANDĚL, Jiří. *Matematická statistika.* 2. vydání. Praha: SNTL, 1985.

[15] RUCKSTUHL, Andreas. *Introduction to Nonlinear Regression.* ZHAW Zürcher Hochschule für Angewandte Wissenschaften, 2010.

[16] GUISAN, Antoine, Thomas C EDWARDS a Trevor HASTIE. Generalized linear and generalized additive models in studies of species distributions: setting the scene. *Ecological Modelling* [online]. 2002, 157(2-3), 89-100 [cit. 2018-01-18]. DOI: 10.1016/S0304-3800(02)00204-1. ISSN 03043800.

[17] FAN, Xingwang a Yuanbo LIU. A Generalized Model for Intersensor NDVI Calibration and Its Comparison With Regression Approaches.*IEEE Transactions on Geoscience and Remote Sensing* [online]. 2017, 55(3), 1842-1852 [cit. 2018-01-18]. DOI: 10.1109/TGRS.2016.2635802.

[18] LANE, Peter W. *Generalized Nonlinear Models.* PRAT, Albert, ed. COMPSTAT. Heidelberg: Physica-Verlag HD, 1996, 1996, s. 331-336.

[19] ZHANG, Guoqiang, B. EDDY PATUWO a Michael Y. HU. Forecasting with artificial neural networks. *International Journal of Forecasting* [online]. 1998, 14(1), 35-62 [cit. 2018-01-18]. DOI: 10.1016/S0169-2070(97)00044-7.

[20] ABBASI, Maryam a Ali EL HANANDEH. Forecasting municipal solid waste generation using artificial intelligence modelling approaches. *Waste Management* [online]. 2016, 56, 13-22 [cit. 2018-01-18]. DOI: 10.1016/j.wasman.2016.05.018.

[21] MWENDA, Amon, Dmitry KUZNETSOV a Silas MIRAU. Time Series Forecasting of Solid Waste Generation in Arusha City - Tanzania. *Mathematical Theory and Modeling.* 2014, 4(8).

[22] GHINEA, Cristina, Elena Niculina DRĂGOI, Elena-Diana COMĂNIŢĂ, Marius GAVRILESCU, Teofil CÂMPEAN, Silvia CURTEANU a Maria GAVRILESCU. Forecasting municipal solid waste generation using prognostic tools and regression analysis. *Journal of Environmental Management.* ECMS, 2016, 2006-05-28, 182(8-9), 80-93. DOI: 10.1016/j.jenvman.2016.07.026.

[23] EMILIA DEN BOER .. (EDS.). *Waste management planning and optimisation: handbook for municipal waste prognosis and sustainability assessment of waste management systems.* Stuttgart: ibidem-Verl, 2005. ISBN 9783898215190.

[24] PAVLAS, Martin, Radovan ŠOMPLÁK, Veronika SMEJKALOVÁ, Vlastimír NEVRLÝ, Lenka ZAVÍRALOVÁ, Jakub KŮDELA a Pavel POPELA. Spatially distributed production data for supply chain models - Forecasting with hazardous waste. *Journal of Cleaner Production.* 2017, (161), 1317-1328.

[25] PAVLAS, Martin. *Justine: tool applied for forecasting in waste management.* [cit. 2018-05-19]. Dostupné z: http://upi.fme.vutbr.cz/media/zalozka_ sekce/pavlas/justine/justine_ description_ r0a.pdf

[26] 383/2001 Sb. Vyhláška Ministerstva životního prostředí o podrobnostech nakládání s odpady.

[27] *Informační systém odpadového hospodářství.* CENIA [online]. [cit. 2018-03-26]. Dostupné z: http://www1.cenia.cz/www/odpady/isoh

[28] 93/2016 Sb. Vyhláška o Katalogu odpadů.

[29] ŠOMPLÁK, Radovan a Martin PAVLAS. Návrh optimální sítě zařízení k nakládání s odpady v rámci celé ČR včetně stanovení potřebných kapacit těchto zařízení ve všech krajích. *Projekty v oblasti odpadového hospodářství k podpoře plnění cílů POH ČR hrazené z PO8 OPŽP 2007 - 2013.* 2015.

[30] ŠOMPLÁK, Radovan, Martin PAVLAS a Veronika SMEJKALOVÁ. Nástroje pro predikci produkce a složení komunálních odpadů. *Waste forum.* 2016, (2), 79-92.

[31] MELOUN, Milan a Jiří MILITKÝ. *Statistická analýza experimentálních dat.* 2. vyd. Praha: Akademie věd České republiky, 2004.

[32] KARPÍŠEK, Zdeněk. *Matematika IV: statistika a pravděpodobnost.* 2. vyd. Brno: Akademické nakladatelství CERM, 2003. ISBN 80-214-2522-9.

[33] ZVÁRA, Karel. *Regrese.* Praha: MATFYZPRESS, 2008. ISBN 987-80-7378-041-8.

[34] DIXON, Juan W. Analysis of extreme values. *The Annals of Mathematical Statistics.* 21(4), 1950, 488-506.

[35] RIAZOSHAMS, A. Hossein, B. Midi HABSHAH, Jr. a C. Mohamad Bakri ADAM. On the outlier Detection in Nonlinear Regression. *International Journal of Mathematical and Computational Sciences.* 2009, 3(12), 1105-1111.

[36] JAVŮREK, M. a I. TAUFER. Testování různých typů reziduí v regresní diagnostice. *CHEMAGAZÍN.* 2011, 21(6), 33-35.

[37] COOK, R. Dennis. a Sanford WEISBERG. *Residuals and influence in regression.* New York: Chapman and Hall, 1982. ISBN 0412242800.

[38] CHATTERJEE, Samprit a Ali S. HADI. *Regression analysis by example.* 4th ed. /. Hoboken, N.J.: Wiley-Interscience, c2006. ISBN 978-0-471-74696-6.

[39] MILITKÝ, Jiří, Karel KVĚTOŇ a Jaroslav CÁP Comparison of some influence measures in nonlinear regression. *Proceedings of the Second International Tampere Conference in Statistics*, 1987, s. 591-602.

[40] EVERITT, Brian. *Cluster analysis.* 5th ed. Chichester, West Sussex, U.K.: Wiley, 2011. Wiley series in probability and statistics. ISBN 9780470977811.

[41] ROUSSEEUW, Peter J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics.* 1987, , 53-65.

[42] KLAPKA, Jindřich, Jiří DVOŘÁK a Pavel POPELA. *Metody operačního výzkumu.* Brno: Vysoké učení technické, 1996. ISBN 80-214-0817-0.

[43] WILLIAMS, H. P. *Model building in mathematical programming.* 5th ed. Hoboken, N.J.: Wiley, 2013. ISBN 978-1-118-44333-0.

[44] BAZARAA, M. S., Hanif D. SHERALI a C. M. SHETTY. *Nonlinear programming: theory and algorithms.* 3rd ed. Hoboken, N.J.: Wiley-Interscience, c2006. ISBN 978-0-471-48600-8.

[45] HAMALA, Milan a Mária TRNOVSKÁ. *Nelineárne programovanie, teória a algoritmy.* Bratislava: EPOS, 2013.

[46] BOYD, Stephen P. a Lieven. VANDENBERGHE. *Convex optimization*. New York: Cambridge University Press, 2004. ISBN 978-0-521-83378-3.

[47] FRANK, Marguerite a Philip WOLFE. An algorithm for quadratic programming. *Naval Research Logistics Quarterly*. 1956, 3(95). DOI: 10.1002/nav.3800030109.

[48] JENSEN, Paul A. a Jonathan F. BARD. *Operations research: models and methods*. Great Britain: Wiley, c2003. ISBN 978-0471380047.

[49] MALAGUTI, Enrico, Silvano MARTELLO a Alberto SANTINI. The traveling salesman problem with pickups, deliveries, and draft limits. 0305-0483. *Omega*. 2018, (74), 50-58.

[50] CHHETRI, Amit S., Darryl MORRELL a Antonia PAPADREOU-SUPPAPPOLA. On the Use of Binary Programming for Sensor Scheduling. *Transactions on signal processing*. 2007, (6).

[51] WOLSEY, Laurence A. *Integer programming*. New York: Wiley, c1998. ISBN 0-471-28366-5.

[52] BELOTTI, Pietro, Christian KIRCHES, Sven LEYFFER, Jeff LINDEROTH, Jim LUEDTKE a Ashutosh MAHAJAN. *Mixed-Integer Nonlinear Optimization*. Illinois, 2012.

[53] NEVRLÝ, Vlastimír, Radovan ŠOMPLÁK, Pavel POPELA, Martin PAVLAS, Ondřej OSIČKA a Jakub KŮDELA. Heuristic challenges for spatially distributed waste production identification problems. *Mendel Journal series*. Brno, 2016, 22(1), 109-116.

[54] Justine – tool applied for forecasting in waste management. Ústav procesního inženýrství, Fakulta strojního inženýrství VUT v Brně [online]. [cit. 2018-01-19]. Dostupné z: http://upi.fme.vutbr.cz/veda-vyzkum/justine.

[55] SZÁSZIOVÁ, Lenka. *Analýza interakcí v odpadovém hospodářství.* Brno: Vysoké učení technické v Brně, Fakulta strojního inženýsrtví, 2017. 93 s. Vedoucí dizertační práce prof. Ing. Petr Stehlík, CSc., dr. h. c.

[56] SMEJKALOVÁ, Veronika, Radovan ŠOMPLÁK, Vlastimír NEVRLÝ a Martin PAVLAS. Heuristic methodology for forecasting of quantities in waste management. *Mendel Journal series.* 2017, 2017(1), 185-192. ISSN 1803-3814.

[57] AGGARWAL, Charu C., Alexander HINNEBURG a Daniel A. KEIM. On the Surprising Behavior of Distance Metrics in High Dimensional Space. *International Conference on Database Theroy*. Springer Berlin Heidelberg, 2001, , 420-434.

[58] TableCurve 2D. https://systatsoftware.com/products/tablecurve-2d (2017). [Online; accessed 4-January- 2018]

[59] ARCHONTOULIS, Sotirios V. a Fernando E. MIGUEZ. Nonlinear Regression Models and Applications in Agricultural Research. *Agronomy Journal* [online]. 2015, 107(2), 786- [cit. 2018-03-05]. DOI: 10.2134/agronj2012.0506. ISSN 0002-1962.

[60] RUCKSTUHL, Andreas. *Introduction to Nonlinear Regression*. ZHAW Zürcher Hochschule für Angewandte Wissenschaften: IDP Institut für Datenanalyse und Prozessdesign, 2010.

[61] RITZ, Christian a JENS CARL STREIBIG. *Nonlinear regression with R*. New York: Springer, 2008. ISBN 9780387096155.

[62] *Zákon č. 185/2001 Sb. Zákon o odpadech a o změně některých dalších zákonů.* In: Sbírka zákonů České republiky, část druhá, § 39 a 40. Dostupný také z /www.platnalegislativa.nsf/

[63] NORDSIECK, Hermann, Matthias SEITZ, Sarah MEYER, Markus HERTEL a Wolfgang ROMMEL. *Separate Waste Collection – Feasible Separation Targets, Market Analysis and Examples of Good Practice*. 28.

[64] BEIGL, P., G. WASSERMANN, F. SCHNEIDER a S. SALHOFER. Forecasting Municipal Solid Waste Generation in Major European Cities. Institute of Waste Management, BOKU.

SEZNAM SYMBOLŮ, VELIČIN A ZKRATEK

CENIA	Česká informační agentura životního prostředí
GAMS	The General Algebraic Modeling System
GLM	Zobecněný lineární model
ISOH	Informační systém odpadového hospodářství
KO	Komunální odpad
KO*	Vybrané složky komunálního odpadu
MINLP	Smíšený celočíselný nelineární problém
MS	Míra separace
MŽP	Ministerstvo životního prostředí
NO	Nebezpečný odpad
OH	Odpadové hospodářství
ORP	Obec s rozšířenou působností
OSTsko	Ostatní složky v SKO
PAP	Papír - separovaný sběr
PAPcel	Celková produkce papíru
PAPsko	Papír v SKO
PL	Plast - separovaný sběr
PLcel	Celková produkce plastu
PLsko	Plast v SKO
PAPcel	Celková produkce papíru
SKL	Sklo - separovaný sběr
SKLcel	Celková produkce skla
SKLsko	Sklo v SKO
SKO	Směsný komunální odpad

ÚPI Ústav procesního inženýrství